

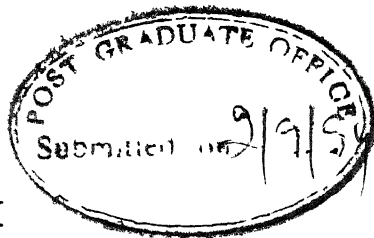
QUEUEING ANALYSIS OF A NON-PREEMPTIVE MMPP/D/1/K PRIORITY SYSTEM FOR APPLICATIONS IN ATM NETWORKS

**A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY**

***By*
B. VENKATARAMANI**

to the

**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
SEPTEMBER, 1994**



CERTIFICATE

It is certified that the work contained in the thesis entitled "Queueing analysis of a non-preemptive MMPP/D/1/K priority system for applications in ATM networks", by B.Venkataramani, has been carried out under our supervision and that this work has not been submitted elsewhere for a degree

Dr K R Srivathsan

Professor,

Dept of Electrical Engg

IIT, Kanpur

Dr Sanjay K Bose

Professor,

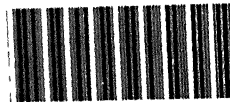
Dept of Electrical Engg

IIT, Kanpur

- 3 JUL 1996
CENTRAL LIBRARY
I I T., KANPUR

Acc. No. A.121800

EC-1004-D-VEN-SUE



A121800

SYNOPSIS

In the Asynchronous Transfer Mode (ATM) networks, incorporation of service priority is desirable for certain real time process control, alarm and network control applications. It has been found that the incorporation of service priority also improves the loss performance of the ATM nodal buffer and reduces the output port contention in an ATM switch with input buffers [1], [2]. In the ATM context, the analysis of a queueing system with multiple service priorities has been reported only for simple input models in the literature, eg for Poisson and Bernoulli models. For example, in [1], Gravey et al assume two priority classes for the traffic arriving at an ATM switch and use a non-preemptive M/D/1 priority model for obtaining the queueing delays for these classes. However these models may not be completely satisfactory for the bursty sources encountered in ATM networks and more complex models, such as Markov Modulated Poisson Process (MMPP) models, may be more desirable. Assuming the MMPP model and a FCFS discipline, several aspects of Call Admission Control (CAC) and nodal buffer design in ATM networks have already been considered in the literature [3]. These approaches do not, however, account for multipriority traffic as results on the queueing model for such traffic are not available for MMPP sources. The contribution made by this thesis is in presenting such a queueing model. Specifically, we present in this thesis the queueing analysis of a multipriority, non-preemptive MMPP/D/1/K system with either infinite or finite buffers for the individual priorities, where the input processes may be MMPP in nature. This analysis is carried out using the matrix analytic approach. Computational results are presented and the approach is verified by comparing these numerical results with those obtained through simulations.

We study a dual priority system first, under the following assumptions -

- 1 The delay sensitive cells and the non-delay sensitive cells arriving at an ATM multiplexer are buffered at two separate queues Q1 and Q2. The arrivals to Q1 and Q2 are from two independent $M(\geq 2)$, $N(\geq 2)$ phase MMPPs
- 2 A single server is shared between the two queues. The cells at Q2 have non-preemptive priority over those at Q1 for receiving service. The priority is incorporated at call level. The server is asynchronous, i.e. a cell arriving when the queue is empty receives service immediately.
- 3 Cells from each priority class require a constant service time of D sec.

We use the words "cell" and "customer" interchangeably as the basic unit of information transferred in an ATM network is a cell.

Using these assumptions the computation of the queue length density (QLD) at Q1 and Q2 at the departure epochs of customers from the respective queues (viz. the probability that the no. of customers in Q1 is equal to 0, 1, at a departure instant from Q1 for $i=1, 2$) is considered first. *For the application of the matrix geometric approach for the present problem, some generalizations of the approach used for the single priority system in [4],[5] are proposed and are as follows*

- 1 The inter-departure time of customers from the low priority queue depends on the phase of the arrival process to the higher priority queues and hence it should be treated as a vector random variable.
- 2 In the priority system, the time when the first customer arrives at an empty queue and the time when the busy period of the server starts need not be identical. In view of this, the busy period of Q1 (for $i=1, 2$) is defined to be the time that elapses since the beginning of the service for the first customer arriving at Q1 and the time when Q1 becomes empty again.

The analysis approach can be summarized as follows. We choose two Semi-Markov chains (SMC), one each corresponding to each priority class with the embedded points as the departure instants of customers from the respective queue. The transition probability matrices of these SMCs are denoted as $Q'(t)$ and $Q''(t)$ respectively. The stationary vectors of $Q'(\infty)$ and $Q''(\infty)$ give the required

QLDs at Q1 and Q2 In view of the non-preemptive priority, the matrices $Q'(\infty)$ and $Q''(\infty)$ are coupled and hence the invariant vectors have to be obtained iteratively

$Q'(t)$ and $Q''(t)$ are expressed in terms of two sets of $MN \times MN$ matrix mass functions $[A'_m(t), B'_m(t)]$ and $[A''_m(t), B''_m(t)]$ respectively Evaluation of these matrix mass functions are considered next and it requires the knowledge of the busy period distribution of Q2, the counting functions associated with the MMPPs to Q1 and Q2 (these functions quantify the probability of there being n arrivals in an interval of time along with the phase transition of the underlying modulating process), the probability that Q1, Q2 is empty at an arbitrary time instant and the QLD of Q2

In order to compute the above matrix functions, evaluation of some of the characteristics of the busy periods of Q1 and Q2 are considered Exploiting the fact that the service time/cell, is constant a recursive procedure for the computation of the busy period distribution is developed Evaluation of the counting functions associated with the MMPPs requires numerical integration of some differential-difference equations for a general service time distribution [4] Exploiting the fact of constant service time/cell an alternate, computationally efficient recursive procedure is proposed for computing these functions using infinite series expansion The convergence problem associated with this series at high traffic rates are overcome by computing them in two stages Expressions for the average number of customers served and the average duration of the busy period at Q1 and Q2 are also obtained

Numerical computation of the QLDs are considered next Towards this end the following issues are considered in detail first

- 1 Computation of the probability of finding zero, one cell at the departure instant of cells from Q1 (for $i=1, 2$) using first passage time arguments
- 2 Computation of the probability of Qi being empty at an arbitrary time t
- 3 Evaluation of the moments of the queue lengths at Q1 and Q2
- 4 Details and the relative advantages of the computation of the QLDs of Q1 and Q2 using (i) Gaussian elimination method (ii) Block Toeplitz inversion method (iii) Recursive procedure

At high traffic loads at Q1, the computational and storage complexity required for the evaluation of the QLDs becomes prohibitively high Under this condition the practical buffer sizes used for Q1 may not be large enough to be treated to be infinite Hence the computation of the QLDs of Q1 and Q2 when Q1

is finite sized is considered next. As an alternative, computation of the QLD of Q2 and the moments of the queue length at Q1 without evaluating the QLD of Q1 is also considered.

It may be noted that the computational and storage complexity required for the evaluation of the QLD of Q1 and Q2 under the priority system is increased by a factor of $O(N^2)$ and $O(M^2)$ over that of the corresponding single priority system. This is because in the priority system the phases of the MMPPs to both Q1 and Q2 need to be tracked at all departure instants. An approximate model which is computationally and storage wise efficient is proposed next. This model keeps track of the phase of only one of the MMPPs at a time. Using this model, evaluation of the QLDs of Q1 and Q2 when the inputs to both Q1 and Q2 are approximated by Poisson processes, is also considered. Finally, some details on the evaluation of the busy period distribution and the QLDs using simulation is considered.

Results on the computation of the QLDs of Q1 and Q2 using the exact model (model I) and approximate model (model II) are presented next for a number of examples and compared with those obtained using simulation. For numerical computations, We assume the traffic to Q1 and Q2 to originate from N_1 Type 1 on/off sources and N_2 Type j on/off sources, respectively ($i=j$ implies identical type of on/off sources to Q1 and Q2). An output link of 150 Mbps and a cell size of 53 bytes are also assumed. The traffic to Q1 and Q2 are approximated by two 2 phase MMPPs using the method proposed in [6]. Knowing the MMPP model parameters, $Q'(\omega)$ and $Q''(\omega)$ are then found and the QLDs are obtained iteratively. Due to resource constraints, the evaluation of the QLD is considered only for cases where the traffic offered to Q2 is less than or equal to 0.35. For cases where the high priority load is greater than 0.35 a finite capacity non-preemptive MMPP/D/1/K priority system is suggested for the evaluation. When the total traffic offered to the server is close to the capacity of the server the computational and storage requirements become high and hence in these cases Q1 buffer size is assumed to be finite. Based on the examples considered the following conclusions are drawn.

- 1 The QLDs of Q1 and Q2 obtained using the exact model agrees well with those obtained using simulation in all the examples considered.
- 2 It appears that model II should not be used if either of the two conditions, (a) or (b), are true - in these cases, the model I is recommended for computations. Otherwise, model II may be preferred due to its simplicity.

- (a) if λ_1''/λ_2'' is significantly larger than 1, where λ_1'' denotes the arrival rate of the MMPP to Q2 in the i th phase and are labelled such that $\lambda_1'' > \lambda_2''$
 - (b) there is a particular phase pair of the two MMPPs which tends to overload the server and this phase pair is fairly likely to arise
- 3 The QLDs of Q1 computed using model II agrees with those of model I at low queue lengths even when the conditions (a) and (b) are true. Because of this the probability of Q1 being empty at an arbitrary time instant computed using model I and II turns out to be essentially the same.
 - 4 The QLD of Q2 computed using all the three methods agree under all conditions.

Finally, the results on the computation of the QLDs of Q1 and Q2 obtained by assuming the traffic to be modelled as Poisson process, are presented for some typical examples. In this case, the QLD of Q1 agrees with that obtained using model II. The QLD of Q2 differs from those of model I and II at higher queue lengths.

Next, the expressions for the distribution of the virtual waiting time of a customer arriving at Q2 and its Laplace Steiltjes Transform are obtained. Using these results, the average queueing delay at Q2 as well as that in Q1 are obtained. Extension of these results for the approximate model as well as the degenerate case of non-preemptive M/D/1 priority system are considered. For the examples considered earlier for the evaluation of the QLDs, the average queueing delays at Q1 and Q2 are computed using the exact as well as approximate models and are found to be in agreement with the results obtained using simulation. For the M/D/1 system, the average queueing delays are computed and are found to be in agreement with the results obtained using an alternate approach given in [1]. The expression for the LST of the virtual waiting time distribution also enables the computation of the percentile of the queueing delays at Q2.

Next, the computation of the QLDs of a non-preemptive MMPP/D/1 dual priority system with finite capacity at Q2 is considered. The busy period distribution (BPD) of the finite capacity system differs from the infinite capacity case as follows:

- 1 Customers arriving when the buffer is full are denied service.
- 2 The maximum number of customers that can be admitted into the system during the service time of a customer depends on the empty space in Q2. Because

of this the distribution of the first passage times are not identical and depends on the state of Q2

As in the infinite case, a recursive procedure for the computation of the BPD of Q2 is developed, using the fact that the service time/customer is constant. Computation of the BPD for the finite capacity case requires considerable computational effort compared to that of the infinite capacity case. This is because, in the present case to compute the BPD of Q2 of capacity N , the BPD of capacity of 1, 2, . . . $N-1$ should be computed first. An indirect method for the efficient computation of the BPD of finite capacity Q2 is proposed. Modifications of the equations of the infinite capacity system for the present case are discussed.

The busy period distribution at Q2 is computed for some examples using both the direct and indirect methods. The latter method is found to require 50% less computation time. The BPD numerically computed is also compared with the simulation results and is found to match well. For some typical examples of traffic from on/off sources, the QLDs at Q1 and Q2 are evaluated using the exact model as well as the approximate model and compared with those obtained using simulation. The conclusions drawn for the infinite capacity is found to be valid for this case as well. Computation of the average queueing delays is also considered for the finite capacity case.

Finally, the computation of the QLDs and the queueing delays of a non-preemptive MMPP/D/1 priority system with more than two priority classes are considered. The traffic from each priority class is assumed to arrive at separate queues and demand the same service time of D sec/customer. The computation of the QLD of a triple priority system is considered first and the extension of this result for higher number of priority classes are indicated. The computation of the average queueing delay also proceeds in a similar fashion.

The major contributions of this thesis are summarized as follows

- 1 Generalization of the matrix analytic approach for the study of a non-preemptive MMPP/D/1/K priority system with either finite or infinite buffer space under the assumption that the traffic from each priority class arrives at separate queues
- 2 Evaluation of the queue length densities, moments of queue lengths and queueing delays at the queues corresponding to each priority class as well

as the percentile of the queueing delay at the high priority queue

- 3 Development of efficient recursive procedures for the computation of the busy period distribution of the server in each of the queues when the buffer size is either infinite or finite
- 4 Development of an efficient procedure for the computation of the counting functions associated with the MMPPs
- 5 Suggestion of two computationally and storage wise efficient approximate models for the priority system and investigation of the range over which they are accurate

REFERENCES:

- 1 A Gravey and G Hebuterne "Mixing time and loss priorities in a single server queue", Proc ITC - 13, Copenhagen, pp 147-152, June- 1991
- 2 Lilly K Jacob and Anurag Kumar, "Comparative performance of scheduling strategies for switching and Multiplexing in a Hub-based ATM network A simulation study" ITC specialist seminar Bangalore, Nov 1993, pp 35-44
- 3 Raif O Onvural, "Asynchronous Transfer mode networks Performance issues", Artech House, Boston, 1994
- 4 M F Neuts, " Structured stochastic matrices of the M/G/1 type and their application", Marcel Dekker, 1989
- 5 V Ramaswami, "The N/G/1 queue and its detailed analysis", Adv Appl Prob Vol 12, pp 222-261, Mar 1980
- 6 H Heffes and D M Lucantoni, " A Markov Modulated characterization of Packetized voice and Data traffic and Related Statistical Multiplexer Performance", IEEE J SAC , No 6, pp 856- 867, Sep 1986

ACKNOWLEDGEMENT

I am deeply indebted to my thesis supervisors Dr Sanjay Bose and Dr K R Srivathsan for all the help and encouragement I received from them. Dr Sanjay Bose introduced me to the world of ATM networks and made available most of the recent literature in this area. Through his brief comments and suggestions, e-mail transactions and electronic editing, he has considerably speedened up my work. In spite of the hectic schedule in the day, Dr K R Srivathsan has spent several long evenings and week ends with me in offering his comments and suggestions on my present work. The help and encouragement received from him during and after my protracted illness has been greatly responsible for my speedy recovery.

I am grateful to Dr P R K Rao, for all the interactions I have had with him during all these years, which developed in me the skills required for carrying out the present work. I thank my teachers Dr V Sinha, Dr. M U Siddiqui, Dr P K Chatterjee and Dr S K Mullick for making the subjects interesting to me.

I am grateful to Mr T Srinivasa Rao for spending his invaluable time with me. The discussions I had with him on developing the simulation routine as well as the methodology used in this thesis cannot be thanked by a few words. I am thankful to Mr Balvinder Singh for helping me to submit the simulation jobs in the image processing laboratory.

I thank Mr Vinod Kumar, Mr Chaturvedi, Mr Abay Karandikar, Mr Deepak Murthy, Mr Venkatesh, Mr Shiv Narayan, Mr Deshpande, Mr Arun Kumar, Mr Chaintanya and Mr Vivek Mudgil for the useful exchange of views.

I am thankful to the engineers Mr Kumar, Mr Whiteford, Ms Padmaja, Mr Sanjeev Singh, Mr Navpreet, Ms Sandhya, Mr A Shukla, Mr Roy, Mr Reddy, Mr Prasanthan and Ms Nandini Mudgil for their help

I would like to thank Mr Bhatnagar, Mr Sadagopan, Mr Murthy, Mr Kousal, Mr Raju, Mr Ajai and Mr Ramnath for their assistance

I am extremely grateful to CSIR for awarding me the Research Associateship for carrying out this work I am thankful to the Principal, Regional Engineering College, Trichirapalli for his encouragement for carrying out this work

There are many names which do not appear in this column but are ringing in my mind These names have been left out here either because they are too obvious or because they have been missed due to my short memory I thank one and all who helped me in my present effort

TABLE OF CONTENTS

CHAPTER 1

Introduction	1
1 1 Introduction	1
1 2 The present PSTN and data networks	1
1 3 Narrowband and broadband ISDN	
1 4 Teletraffic engineering	6
1 5 Statement of the problem	8
1 6 Organization of the thesis	8
References	12

CHAPTER 2

Review of the previous work and an overview of the problem	13
2 1 Introduction	
2 2 Introduction to ATM networks	14
2 2 1 Principles of ATM	14
2 2 2 Specific characteristics of ATM networks	15
2 2 4 Layered architecture for ATM networks	18
2 3 Some of the design issues in ATM networks	18
2 3 1 Call admission control	19
2 3 2 Traffic parameter control (policing function)	21
2 3 3 Congestion control in ATM networks	22
2 3 5 Incorporation of priority in ATM networks	24
Space priority mechanisms	24
Incorporation of time priority	25
2 3 6 Traffic models for ATM sources	26
2 4 Review of the previous work	29

2 5	An overview of the problem	38
	References	41

CHAPTER 3

	Embedded semi-Markov sequences and transition probability matrices of the high and low priority queues	50
3 1	Introduction	50
3 2	Embedded semi - Markov sequences of Q1 and Q2	53
3 3	Structure of the transition probability matrices of Q1 and Q2	55
3 4	Computation of $A_m''(t)$ and $U_k''(t)$	64
3 5	Computation of $A_m'(t)$ and $U_k'(t)$	66
3 6	Transforms of the elements of $Q'(\cdot)$ and $Q''(\cdot)$	72
	References	77

CHAPTER 4

	Characteristics of busy periods of Q1 and Q2	78
4 1	Introduction	78
4 2	First passage times and busy periods of $Q'(\cdot)$ and $Q''(\cdot)$	
4 3	The busy period distribution of Q1 and Q2	85
4 4	Computation of $P''(m,D)$ and $P'(m,nD)$	89
4 5	Average durations of the busy periods and the number of cells served during the busy periods of Q1 and Q2	93
4 6	Average duration of the busy cycles of Q1 and Q2	100
	References	105

CHAPTER 5

Evaluation of the queue length densities of Q1 and Q2	107
5 1 Introduction	107
5 2 Expressions for x'_0 and x''_0	107
5 3 Computation of x''_1 and x'_1	109
5 4 Moments of the queue lengths of Q1 and Q2	111
5 5 Computation of y'_0 and y''_0	116
5 6 Stationary queue length distribution of Q1 and Q2	122
Gaussian elimination method	124
Block Toeplitz matrix inversion method	126
Recursive procedure	129
Evaluation of the QLDs for some special cases	130
1 Q1 finite sized and Q2 infinite sized	130
2 Evaluation of the QLD of Q2 and the moments of the queue lengths of Q1	130
5 7 An approximate model for the time priority system	130
5 8 Some assumptions and approximations for numerical computations	137
Computation of the model parameters of MMPP	137
Truncation of the infinite sums for the evaluation of $Q'(\infty)$	141
Evaluation of numerical integrals	141
Computation of the exponential matrices $e^{Q \cdot nD}$	142
5 9 Simulation of the priority system	143
Simulation with on/off source model	147
Simulation of busy period distribution with MMPP model	150
Model for the computation of the QLDs and queueing delays at Q1 and Q2	154

	Computation of the confidence intervals for the probabilities	161
5 10	Numerical results	163
	References	179

CHAPTER 6

	Queueing delay of a non-preemptive MMPP/D/1 dual priority system	209
6 1	Introduction	209
6 2	Virtual waiting time distribution of a high priority cell	210
6 3	LST of the virtual waiting time distribution of Q2 cells	217
6 4	Virtual waiting time distribution and its LST using the approximate model	224
6 5	Average queueing delay of Q2	228
6 6	Average queueing delay at Q1	236
6 7	Virtual waiting time of M/D/1 queues with non-preemptive priority	238
6 8	Numerical Results	240
	Appendix (6 A)	242
	Appendix (6 B)	245
	Appendix (6 C)	248
	References	251

CHAPTER 7

	Computation of the queue length densities of a non-preemptive MMPP/D/1/K dual priority system	254
7 1	Introduction	254
7 2	Embedded semi-Markov sequence and state transition probability matrices of Q1 and Q2	255
7 3	Characteristics of the busy period of Q2	262

	Probability generating function of $G^{(N,k)}(l,t)$	262
	Computation of the busy period distribution of Q2	265
	Computation of the average duration and the number of customers served during the busy periods of Q2	271
7 4	Computation of the stationary queue length densities at Q1 and Q2	275
	Computation of x_0'' and y_0''	276
	Computation of the QLDs using the approximate model	277
	Numerical Results	277
	References	282
CHAPTER 8		
	Extension for three or more priority classes	289
8 1	Introduction	289
8 2	Evaluation of the QLDs of a non-preemptive MMPP/D/1 triple priority system	289
8 3	Extension for more than three priority classes	295
8 4	Computation of the average queueing delays of the multi-priority system	296
	Appendix (8 A)	297
	References	302
CHAPTER 9		
Conclusions		
9 1	Introduction	303
9 2	Application of the matrix analytic approach for the priority system	303

9 3	Some observations and conclusions drawn in the evaluation of the QLDs of the dual priority system	304
9 4	Evaluation of the queueing delays	308
9 5	Finite capacity dual priority system	309
9 6	Extension for three or more priority classes	310
9 7	Suggestions for further work	310

LIST OF FIGURES

Fig 1 1	Basic functional elements of a public ISDN	4
Fig 1 2	Approximate ATM traffic performance requirements	6
Fig 2 1	Synchronous time division multiplexing	15
Fig 2 2	Statistical time division multiplexing	15
Fig 2 3	Organization of the cell	17
Fig 2 4	Layered architecture for ATM networks	18
Fig 2 5	ATM multiplexer and the buffer	19
Fig 3 1	Model for the time priority mechanism	51
Fig 3 2	Arrivals and phase transitions at Q1 during the IDT of Q1 given that Q1 empty at the last departure instant	63
Fig 3 3	Arrivals and phase transitions at Q2 between adjacent Q1 service beginning epochs given Q1 not empty at the previous departure instant	68
Fig 3 4	Time of occurrence of various events in the idle period of Q1 and the phase of the MMPP to Q1 at these instants	70
Fig 5 1	Flow control for the next-event time advance approach	182
Fig 5 2 a	The events types for the evaluation of the BPD through simulation using on/off source model	148
Fig 5 2 b	The events types for the evaluation of the BPD through simulation using on/off source model	151
Fig 5 3	The events types for the evaluation of the QLDs through simulation using on/off source model	156
Fig 5 4	P M F of busy period length of MMPP/D/1 queue computed and simulated for $N_s = 45$ ($\rho = 0.10$)	183
Fig 5 5	P M F of busy period length of MMPP/D/1 queue computed and simulated for $N_s = 90$ ($\rho = 0.21$)	183

Fig 5 6	P M F of busy period length of MMPP/D/1 queue computed and simulated for $N_s = 150$ ($\rho = 0.35$)	184
Fig 5 7	Probability mass function of the busy period length of MMPP/D/1 queue computed for $\rho < 0.5$	185
Fig 5 8	Probability mass function of the busy period length of MMPP/D/1 queue computed for $\rho > 0.5$	186
Fig 5 9	QLD of Q1, the low priority queue computed and simulated for (300,45) type 1 sources ($\rho = 0.7, 0.1$)	187
Fig 5 10	QLD of Q2, the high priority queue computed and simulated for (300,45) type 1 sources ($\rho = 0.7, 0.1$)	187
Fig 5 11	QLD of Q1, the low priority queue computed and simulated for (300,45) type 2 sources ($\rho = 0.7, 0.1$)	188
Fig 5 12	QLD of Q2, the high priority queue computed and simulated for (300,45) type 2 sources ($\rho = 0.7, 0.1$)	188
Fig 5 13	QLD of Q1, the low priority queue computed and simulated for (300,45) type 3 sources ($\rho = 0.7, 0.1$)	189
Fig 5 14	QLD of Q2, the high priority queue computed and simulated for (300,45) type 3 sources ($\rho = 0.7, 0.1$)	189
Fig 5 15	QLD of Q1, the low priority queue computed and simulated for (105,15) type 4 sources ($\rho = 0.7, 0.1$)	190
Fig 5 16	QLD of Q2, the high priority queue computed and simulated for (105,45) type 4 sources ($\rho = 0.7, 0.1$)	190
Fig 5 17	QLD of Q1, the low priority queue computed and simulated for (105,15) type 5 sources ($\rho = 0.7, 0.1$)	191
Fig 5 18	QLD of Q2, the high priority queue computed and simulated for (105,15) type 5 sources ($\rho = 0.7, 0.1$)	191
Fig 5 19	QLD of Q1, the low priority queue computed and	

	simulated for (150,90) type 1 sources ($\rho = 0.35, 0.2$)	192
Fig 5 20	QLD of Q2, the high priority queue computed and simulated for (150,15) type 1 sources ($\rho = 0.35, 0.2$)	192
Fig 5 21	QLD of Q1, the low priority queue computed and simulated for A (300 Type 1, Bursty traffic) at Q1, Q2	193
Fig 5 22	QLD of Q2, the high priority queue computed and simulated for A (300 type 1, Bursty traffic) at Q1, Q2	193
Fig 5 23	QLD of the low priority queue computed and simulated for (300 Type 1,16 type 4) sources (av $\rho = 0.70, 0.1$)	194
Fig 5 24	QLD of the high priority queue computed and simulated for (300 type 1,Bursty traffic) sources ($\rho = 0.70,0.1$)	194
Fig 5 25	QLD of the low priority queue computed and simulated for (350,45) type 1 sources Q1 size=400($\rho = 0.81, 0.1$)	195
Fig 5 26	QLD of the low priority queue computed and simulated for (350,45) type 1 sources Q1 size=100($\rho = 0.81, 0.1$)	195
Fig 5 27	QLD of the high priority queue computed and simulated for (350,45) type 1 sources & Q1 size=INF($\rho = 0.81,0.1$)	196
Fig 5 28	QLD of the high priority queue computed and simulated for (350,45) type 1 sources & Q1 size=475($\rho = 0.81,0.1$)	196
Fig 5 29	QLD of the low priority queue computed and simulated for (360,45) type 1 sources ($\rho = 0.84, 0.1$)	197
Fig 5 30	QLD of the low priority queue computed and simulated for (360,45) type 1 sources ($\rho = 0.84, 0.1$)	197
Fig 5 31	QLD of the low priority queue computed and simulated for (300,90) type 1 sources & Q1 size=INF ($\rho=0.70, 0.2$)	198
Fig 5 32	QLD of the low priority queue computed and simulated for (300,45) type 1 sources & Q1 size=100($\rho=0.70, 0.2$)	198

Fig 5 33	QLD of the high priority queue computed and simulated for (300,90) type 1 sources & Q1 size=INF($\rho = 0.7, 0.1$)	199
Fig 5 34	QLD of the high priority queue computed and simulated for (320,90) type 1 sources & Q1 size=475($\rho = 0.74, 0.1$)	200
Fig 5 35	QLD of the low priority queue computed and simulated for (320,90) type 1 sources & Q1 size=475($\rho = 0.75, 0.2$)	200
Fig 5 36	QLD of the low priority queue computed and simulated for (300,90) type 2 sources & Q1 size=INF($\rho = 0.75, 0.2$)	201
Fig 5 37	QLD of the high priority queue computed and simulated for (300,90) type 2 sources & Q1 size=INF($\rho = 0.56, 0.31$)	201
Fig 5 38	QLD of the low priority queue computed and simulated for (300,90) type 3 sources & Q1 size=INF($\rho = 0.70, 0.2$)	202
Fig 5 39	QLD of the high priority queue computed and simulated for (300,90) type 3 sources & Q1 size=INF($\rho = 0.7, 0.21$)	202
Fig 5 40	QLD of the low priority queue computed and simulated for (240,135) type 1 sources & Q1 size=INF($\rho = 0.56, 0.3$)	203
Fig 5 41	QLD of the high priority queue computed and simulated for (240,135) type 1 sources & Q1 size=INF($\rho = 0.56, 0.31$)	203
Fig 5 42	Variation of P (low priority queue empty) with traffic offered in Q1 and Q2 in an MMPP/D/1 queue	204
Fig 5 43	QLD of the low priority queue computed using models I, II and Poisson for (240,135) type 1 sources($\rho=0.56, 0.31$)	205
Fig 5 44	QLD of the high priority queue computed using models I, II and Poisson for (240,135) type 1 sources($\rho=0.56, 0.31$)	206
Fig 5 45	Comparison of the QLDs of Q1 and Q2 with FCFS queues with dedicated servers for traffics from (300,45) type 1 sources	207
Fig 5 46	Comparison of the QLDs of Q1 and Q2 with FCFS queues	

	with dedicated servers for traffics from (300,45) type 1 sources	208
Fig 6 1	Mean delay for Q2 cells in an MMPP/D/1 priority system obtained using simulation and computation model I	252
Fig 6 2	Mean delay for Q1 cells in an MMPP/D/1 priority system obtained using simulation and computation model I and simulation	252
Fig 6 3	Mean delay for Q2 cells in an MMPP/D/1 priority system computed using models I(exact) and II(appr)	253
Fig 6 4	Mean delay for Q1 cells in an M/D/1 priority system computed using MMPP/D/1 results and alternate approach	253
Fig 7 1	P M F of the busy period length of MMPP/D/1/K queue obtained through computation and simulation for $\rho=0.42$	283
Fig 7 2	P M F of the busy period length of MMPP/D/1/K queue obtained through computation and simulation for $\rho=0.56$	283
Fig 7 3	QLD of Q1, the low priority queue computed and simulated for (180,180) type 1 sources and Q2 size=25 ($\rho = 0.42, 0.42$)	284
Fig 7 4	QLD of Q1, the low priority queue computed and simulated for (180,180) type 1 sources and Q2 size=25 ($\rho = 0.42, 0.42$)	284
Fig 7 5	QLD of Q1, the low priority queue computed and simulated for (150,240) type 1 sources and Q2 size=25 ($\rho = 0.35, 0.56$)	285
Fig 7 6	QLD of Q1, the high priority queue computed and	

	simulated for (150,240) type 1 sources and Q2 size=25 ($\rho = 0.35, 0.56$)	285
Fig 7 7	QLD of Q1, the low priority queue computed and simulated for (225,180) type 1 sources and Q2 size=25 ($\rho = 0.52, 0.42$)	286
Fig 7 8	QLD of Q1, the high priority queue computed and simulated for (225,180) type 1 sources and Q2 size=25 ($\rho = 0.52, 0.42$)	286
Fig 7 9	QLD of Q1, the low priority queue computed and simulated for (165,240) type 1 sources and Q2 size=25 ($\rho = 0.385, 0.56$)	287
Fig 7 10	QLD of Q1, the high priority queue computed and simulated for (165,240) type 1 sources and Q2 size=25 ($\rho = 0.385, 0.56$)	287
Fig 7 11	Variation of P (low priority queue empty) with traffic offered in Q1 and Q2 in an MMPP/D/1/K queue	288
Fig 7 12	Mean delay for Q2 cells in an MMPP/D/1/K priority system obtained using model I, II and simulation	288

LIST OF TABLES

Table 2 1	Queueing delay and transmission time (in μsec)	16
Table 2 2	Propagation delay	16
Table 2 3	Packet loss probability as a function of buffer size	17
Table 2 4	Parameters of the generalized on/off source model for some typical sources	27
Table 5 1	Characteristics of the different constant bit rate on/off sources	164

LIST OF ABBREVIATIONS

ATM	Asynchronous Transfer Mode
MMPP	Markov Modulated Poisson Process
PSTN	Public Switched Telephone Network
FCFS	First Come First Served
CAC	Call Admission Control
QLD	Queue length density
SMC	Semi-Markov chain
BP	Busy period
BPD	Busy period distribution
N-ISDN	Narrowband Integrated Services Digital Networks
B-ISDN	Broadband Integrated Services Digital Networks
QOS	Quality of service
LAN	Local Area Network
TDM	Time Division Multiplexing
STM	Synchronous Time Division Multiplexing
MUX	Multiplexer
SONET	Synchronous Optical Fibre Network
VP	Virtual path
LB	Leaky bucket
LSB	Least Significant byte
CLP	Cell loss priority
FCI	Forward congestion indicator
CBR	Constant bit rate
VBR	Variable bit rate
IPP	Interrupted Poisson Process
AR	Autoregressive Models

PPP	Partial preemptive priority
CLAD	Cell Assembler Disassembler
HOL	Head of the Line
LDOLL	Low delay or low loss
LL	Low loss
LD	Low delay
EDD	Earliest due date
HOL-PJ	Head of line with priority jumps
WRR	Weighted Round Robin
SBBP	Switched Batch Bernoulli process
IDT	Inter departure time
RST	Residual service time
n e	not empty
ABP	Additional busy period
LST	Laplace Steiltjes Transform
MRT	Mean Recurrence Time
MRP	Markov Renewal process
BUTT	Block upper triangular Toeplitz matrix
i i d	independent and identically distributed
PMF	Probability mass function
KRT	Key renewal theorem

LIST OF SYMBOLS

Q_1, Q_2	Low, High priority queues
Q	Composite queue fed by both priority traffic and served on FCFS basis
$MMPP_1$	MMPP model used for the traffic arriving at Q_1
$MMPP_2$	MMPP model used for the traffic arriving at Q_2
$MMPP_1$	MMPP model used for the traffic arriving at Q_1 and for keeping track of the phases of both $MMPP_1$ and $MMPP_2$
$MMPP_2$	MMPP model used for the traffic arriving at Q_2 and for keeping track of the phases of both $MMPP_1$ and $MMPP_2$
(Q^*, Q^{**})	Infinitesimal generator matrices of $MMPP_1, MMPP_2$ to Q_1 and Q_2
\underline{Q}^*	Infinitesimal generator matrix of the composite process obtained by superposing $MMPP_1$ and $MMPP_2$
(Λ', Λ'')	Arrival rate matrices of MMPPs to Q_1 and Q_2
$(\underline{\Lambda}', \underline{\Lambda}'', \underline{\Lambda})$	Arrival rates at Q_1, Q_2 and Q corresponding to the different phases of the composite process
(M, N)	No of phases of $MMPP_1, MMPP_2$
$X(t), \underline{J}(t)$	No of cells in Q , Phase of the composite process at time t
$X'(t), X''(t)$	No of cells in Q_1, Q_2 at time t
$P[]$	the probability of $[]$
$ $	given that
$P'(n, t)$	$MN \times MN$ matrices whose $(i, j)^{th}$ elements denote $P[N'(t)=n, \underline{J}(t)=j \mid N'(0)=0, \underline{J}(0)=1]$
$N'(t)$	No of arrivals at Q_1 in $(0, t]$
$P''(n, t)$	$MN \times MN$ matrices whose $(i, j)^{th}$ elements denote $P[N''(t)=n, \underline{J}(t)=j \mid N''(0)=0, \underline{J}(0)=1]$
$N''(t)$	No of arrivals at Q_2 in $(0, t]$

(τ'_n, τ''_n)	Departure epochs of cells from Q1 and Q2
X'_n, X''_n	No of cells in Q1, Q2 at τ'_n, τ''_n
$J^{(1)'}_n, J^{(2)'}_n$	Phase of MMPP1, MMPP2 at τ'_n
$J^{(1)''}_n, J^{(2)''}_n$	Phase of MMPP1, MMPP2 at τ''_n
$\underline{J}(t)$	Phase of the composite process at time t
$\underline{J}'_n, \underline{J}''_n$	Phase of the composite process at τ'_n, τ''_n
$Q'(t)$	Transition probability matrix of the SMC pertaining to Q1
$A'_m(t)$	MN x MN matrices whose $(i, j)^{th}$ elements denote P[Given that a cell departed from Q1 at time 0, leaving at least one cell in Q1 and the arrival process MMPP 1 in phase 1, the next departure occurs at no later than time t with MMPP 1 in phase j, and in the intervening period there were m arrivals]
$B'_m(t)$	MN x MN matrices whose $(i, j)^{th}$ elements denote P[Given that a cell departed from Q1 at time 0 leaving Q1 empty and the arrival process MMPP 1 in phase 1, the next departure occurs at time no later than t with MMPP 1 in phase j, leaving m cells in Q1]
$U'_k(t)$	MN x MN matrices whose $(i, j)^{th}$ elements denote P[BP of Q1 starts at or before time t, k cells arrive at Q1 in (0, t], $\underline{J}(t) = j X'(0)=0, \underline{J}(0)=1]$
$H'_m(t)$	MN x MN diagonal matrix whose i^{th} diagonal element denotes $P[(\tau'_n - \tau'_{n-1}) \leq t \underline{J}'_{n-1} = 1 \text{ and } X'_{n-1} > 0]$
$Q''(t)$	Transition probability matrix of the SMC pertaining to Q2
$A''_m(t)$	MN x MN matrices whose $(i, j)^{th}$ elements denote P[Given that a cell departed from Q2 at time 0, leaving at least one cell in Q2 and the arrival process MMPP 2 in phase 1, the next departure occurs at no later than time t with MMPP 2 in phase j, and in the intervening period there were m arrivals]

$B_m''(t)$	MN x MN matrices whose $(i,j)^{th}$ elements denote $P[\text{Given that a cell departed from Q2 at time 0, leaving Q2 empty and the arrival process MMPP } \underline{2} \text{ in phase 1, the next departure occurs at time } \leq t \text{ with MMPP } \underline{2} \text{ in phase } j, \text{ leaving } m \text{ cells in Q2}]$
$U_k''(t)$	MN x MN matrices whose $(i,j)^{th}$ elements denote $P[\text{Busy period of Q2 starts at or before time } t, k \text{ cells arrive at Q2 in } (0,t], \underline{J}(t) = j X''(0)=0, \underline{J}(0)=1]$
$H''(t)$	MN x MN diagonal matrices whose i^{th} diagonal element denotes $P[(\tau_n'' - \tau_{n-1}'') \leq t, \underline{J}_{n-1}'' = 1 \text{ and } X_{n-1}'' > 0] \delta_{1j}$
x', x''	Invariant probability vector of $Q'(\omega), Q''(\omega)$
x_m'	1 x MN vector whose i^{th} element denotes $P[X_n' = m, \underline{J}_n' = 1]$
y_0'	1 x MN vector whose i^{th} element denotes $P[X'(t) = 0, \underline{J}(t) = 1]$ at any time t
x_m''	1 x MN vector whose i^{th} element denotes $P[X_n'' = m, \underline{J}_n'' = 1]$
y_0''	1 x MN vector whose i^{th} element denotes $P[X''(t) = 0, \underline{J}(t) = 1]$ at any time t
p_0', p_0''	$P[Q1 \text{ empty at time } t], P[Q2 \text{ empty at time } t]$
$U_k''(t)$	MN x MN matrices whose $(i,j)^{th}$ elements denote the $P[\text{BP of Q2 starts at or before } t, Q1 \text{ non-empty when the 1st cell arrives at Q2, } X''(t)=k, \underline{J}(t) = j X''(0)=0, \underline{J}(0)=1,]$
τ_n	Departure epoch of cells from Q
X_n, J_n	No. of cells in Q1, Phase of MMPP at τ_n
$Q(t)$	Transition probability matrix of the SMC pertaining to Q
x	Invariant probability vector of $Q(\omega)$
x_m	1 x MN vector whose i^{th} element denotes $P[X_n = m, \underline{J}_n = 1]$
y_0	1 x MN vector whose i^{th} element denotes $P[X(t) = 0, \underline{J}(t) = 1]$ at any time t

$U_k(t)$	MN x MN matrices whose $(i,j)^{th}$ elements denote the P[the first batch of customers of size k arrive at or before time t, $J(t) = j$ the queue is empty at time 0 and $J(0) = 1$]
$P(n,t)$	MN x MN matrices whose $(i,j)^{th}$ elements denote the P[$N(t)=n, J(t)=j$ $N(0)=0, J(0)=1$]
$N(t)$	No. of arrivals at Q in $(0,t]$
$u(t)$	unit step function
$\delta'(t), \delta_{lk}$	Dirac, Kronecker delta
e	MN x 1 unit vector
I_N	N x N identity matrix
\otimes	Kronecker product of matrices
$\underline{G}^{(m)}(t)$	MN x MN matrices whose $(i,j)^{th}$ elements denote the P[BP of Q2 ends at or before time t, $\underline{J}(t)=j$ it starts at time 0 with $X''(0) = m$ and $\underline{J}(0) = 1$]
$\underline{G}_k^{(m)}$	MN x MN matrices whose $(i,j)^{th}$ elements denote the P[BP of Q2 ends at time $t = kD$ sec, $\underline{J}(t)=j$ it starts at time 0 with $X''(0) = m$ and $\underline{J}(0) = 1$]
$C(k)$	MN x MN diagonal matrices whose i^{th} diagonal is equal to P[BP of Q2 = kD sec $J''_{n-1}=1, X'_{n-1}>0$]
$c(k)$	MN x 1 vectors whose i^{th} element denotes the joint probability that the BP of Q2 intervening between successive departures from Q1 is of duration kD sec and starts in phase 1
$F(k)$	MN x MN diagonal matrices whose i^{th} diagonal element denotes P[ABP = kD sec ABP starts at time 0 with MMPP $\underline{2}$ in phase 1]
$f(\iota)$	MN x 1 vectors whose i^{th} element denotes the joint probability that the ABP of Q2 is of duration ιD and starts in phase 1
$\mathcal{P}'(z,t)$	the z-transform of $P'(m,t)$

$\mathcal{P}''(z,t)$	the z-transform of $\mathbf{P}''(m,t)$
$\tilde{\mathbf{A}}_m''(s), \tilde{\mathbf{U}}_k''(s)$	the LST of $\mathbf{A}_m''(t), \mathbf{U}_k''(t)$,
$\tilde{\mathbf{A}}_m'(s), \tilde{\mathbf{U}}_k'(s)$	the LST of $\mathbf{A}_m'(t), \mathbf{U}_k'(t)$,
$\tilde{\mathbf{A}}''(z,s)$	the z transform of the LST of $\mathbf{A}_m''(t)$
$\tilde{\mathbf{B}}''(z,s)$	the z transform of the LST of $\mathbf{B}_m''(t)$
$\tilde{\mathbf{U}}''(z,s)$	the z transform of the LST of $\mathbf{U}_k''(t)$
$\tilde{\mathbf{A}}'(z,s)$	the z transform of the LST of $\mathbf{A}_m'(t)$
$\tilde{\mathbf{U}}'(z,s)$	the z transform of the LST of $\mathbf{U}_k'(t)$
$\tilde{\mathbf{B}}'(z,s)$	the z transform of the LST of $\mathbf{B}_m'(t)$
$\mathbf{G}''^{(1)}(k,t)$	the MNxMN matrices whose (j,j') th element denotes the probability, given that the semi-Markov process $\mathbf{Q}''(\cdot)$ starts in the state $(1,j)$, it reaches the state $(0,j')$ for the first time after k transitions and the time of such a first passage is at most t
$\tilde{\mathbf{G}}''(z,s)$	the z transform of the LST of $\mathbf{G}''^{(1)}(k,t)$
$\tilde{\mathbf{G}}''^{(1)}(z,s)$	the double transform of $\mathbf{G}''^{(1)}(k,t)$
\mathbf{G}''	MNxMN matrix whose (i,j) th element denotes $P[\text{BP of } Q2 \text{ ends in phase } j \mid \text{it started with MMPP } \underline{2} \text{ in phase } 1]$
\mathbf{g}''	invariant probability vector of \mathbf{G}''
$\mathbf{G}'^{(1)}(k,t)$	the MNxMN matrices whose (j,j') th element denotes the probability, given that the semi-Markov process $\mathbf{Q}'(\cdot)$ starts in the state $(1,j)$, it reaches the state $(0,j')$ for the first time after k transitions and the time of such a first passage is at most t
$\mathbf{G}_n'^{(k)}$	MNxMN matrices whose (i,j) th element denote $P[\text{BP of } Q1 \text{ is of duration } nD \text{ sec and ends with MMPP } \underline{1} \text{ in phase } j \text{ given it started at time } 0 \text{ with } k \text{ customers in } Q1 \text{ and with MMPP } \underline{1} \text{ in phase } 1]$

G'	MN×MN matrix whose $(i,j)^{th}$ element denotes $P[BP \text{ of } Q1 \text{ ends in phase } j \text{it started with MMPP } \underline{1} \text{ in phase } i]$
ρ''	the average traffic offered at Q2
π''	the 1×MN invariant probability vector of $\tilde{A}''(1,0)$
β''	MN×1 vector whose i^{th} element, denotes the average number of customers arriving at Q2 during the IDT of customers from Q2 given that at the previous departure instant Q2 is non-empty and the MMPP $\underline{2}$ is in phase i
g'	invariant probability vector of G'
ρ'	the average traffic offered at Q1
π'	the 1×MN invariant probability vector of $\tilde{A}'(1,0)$
β'	MN×1 vector whose i^{th} element, denotes the average number of customers arriving at Q1 during the IDT of customers from Q1 given that at the previous departure instant Q1 is non-empty and the MMPP $\underline{1}$ is in phase i
$Y(k)$	a matrix array whose i^{th} element gives the coefficient of ξ^i in $(G_1''\xi + G_2''\xi^2 + G_3''\xi^3 + \dots)^k$
$V(n)$	a matrix array whose i^{th} element gives the coefficient of z^{i-1} in $\{[\underline{A}''(z-1) + \underline{Q}^*]D\}^n$
$L''(k,t)$	MN×MN matrices whose $(i,j)^{th}$ elements denote $P[\text{the first BP of } Q2 \leq t, \text{ consists of } k \text{ services and the phase}$ $\text{of MMPP } \underline{2} \text{ is } j \text{ when the BP ends} X''(0)=0 \text{ and } \underline{J}(0)=1]$
$\tilde{L}''(z,s)$	the double transform of $L''(k,t)$
μ''	a vector whose j^{th} element denotes the expected first passage time from $(i+1,j)$ to (i,j) for $i \geq 0$ in the SMP $Q''(\cdot)$
$\tilde{\mu}''$	a vector whose j^{th} element denotes the no of service comple- tions during the first passage time from $(i+1,j)$ to (i,j)

$\mu^{(1)''}$	the average IDT of cells from Q2 given that the previous departure left Q2 non-empty
$\tilde{\mu}_1''$	MNxl vectors whose i^{th} element denotes the mean number of customers served during the first busy period of Q2 given that $X''(0)=0$ and $\underline{J}''(0)=\underline{j}$
$\hat{\mu}''$	MNxl vector whose i th elements denote the average length of the busy cycle of Q2 given that the BP started in phase i
μ_1'', μ_2''	MNxl vectors whose i th elements denote the average length of the vacation interval and the BP of Q2 given that the BP started in phase i
μ'	a vector whose j^{th} element denotes the expected first passage time from $(i+1, j)$ to (i, j) for $i \geq 0$ in the SMP $Q'(\cdot)$
$\tilde{\mu}'$	a vector whose j^{th} element denotes the no of service completions during the first passage of $Q'(\cdot)$ from $(i+1, j)$ to (i, j)
$\mu^{(1)'}$	the average IDT of cells from Q1 given that the previous departure left Q1 non-empty
$\tilde{\mu}_1'$	MNxl vectors whose i^{th} element denotes the mean number of customers served during the first busy period of Q1 given that $X'(0)=0$ and $\underline{J}'(0)=\underline{j}$
$\hat{\mu}'$	MNxl vector whose i^{th} elements denote the average length of the busy cycle of Q2 given that the BP started in phase i
μ_1', μ_2'	MNxl vectors whose i^{th} elements denote the average length of the vacation interval and the BP of Q1 given that the BP started in phase i
$K_0''(n, t)$	MN x MN matrix whose $(i, j)^{\text{th}}$ element gives the joint probability that the busy cycle of Q2 is of length $\leq t$, consists of n customer services and ends with the MMPP $\underline{2}$ phase as j given that at time 0 the phase is i

$\tilde{K}_0''(z,s)$	the double transform of $K_0''(n,t)$
k_0'' , k_0'	the stationary vector of $\tilde{K}_0''(1,0)$ $\tilde{K}_0'(1,0)$
$K_1''(n,t)$	MN \times MN matrix whose $(1,j)^{th}$ element denotes the probability that starting in $(1,1)$, the MRP $Q''()$ returns for the first time to the set 1 in exactly n steps at or before time t and that the phase of MMPP $\underline{2}$ at the epoch of such a first return is j
$\tilde{K}_1''(z,s)$	the double transform of $K_1''(n,t)$
k_1''	the stationary vector of $\tilde{K}_1''(1,0)$
\tilde{k}_1''	the mean number of steps in a first passage from 1 to itself
$X''(z)$	the z transform of x_n''
B_m'' , A_m''	$\tilde{B}_m''(s)$, $\tilde{A}_m''(s)$ evaluated at $s=0$
$X^{(n)}$, $A^{(n)}$	the n^{th} moments of $X''(z)$, $\tilde{A}''(z,0)$ w r t z evaluated at $z=1$
$\mathcal{U}^{(n)}$	the n^{th} moment of $\mathcal{U}(z)$ w r t z evaluated at $z=1$
X , A , \mathcal{U}	$X''(z)$, $\tilde{A}''(z,0)$ and $\mathcal{U}(z)$ evaluated at $z=1$
$\mathcal{P}^{(n)}$	the n^{th} moment of $\mathcal{P}(z,t)$
$U^{(n)}$	the n^{th} moment of $\tilde{U}''(z,0)$
$\tilde{U}^{(n)}$	the n^{th} moment of $\tilde{U}'(z,0)$
$\Phi_{ml}''(t)$	the expected number of visits of the semi-Markov process $Q''()$ to the state (m,ℓ) in the interval $(0,t]$ given that it started from the state $(1,j')$ at time 0
$y''(0,j,t)$	Probability that the SMP $Q''()$ visits the state $(0,j)$ at time t given that at time 0 the state was $(1,j')$ for $1 \geq 0$, $1 \leq j' \leq MN$
δ_T	the mean sojourn time of the process $Q''()$ averaged over all possible states
δ_1	MN \times 1 vector whose j^{th} element, $\delta(1,j)$, denotes the mean sojourn time of the process $Q''()$ in the state $(0,j)$
$m''(1,j)$	the MRT of the state $(1,j)$ of $Q''()$

ℓ_{1j}	the MRT of the state $(1,j)$ of the SMC $Q''(\infty)$
ξ^{**}	the inverse of the mean sojourn time of $Q''(\cdot)$
$m'(1,j)$	the MRT of the state $(1,j)$ of $Q'(\cdot)$
ξ^{**}	the inverse of the mean sojourn time of $Q'(\cdot)$
N_2	the number of sources generating the traffic to Q_2 ,
N_1	the number of sources generating the traffic to Q_1
Model I	Exact model
Model II	Approximate model
$v''(t)$	The virtual waiting time of a cell arriving at Q_2 at time t
$H(\sigma), H^{<n>}(\sigma)$	the c d f of the service time and the n fold convolution of $H(\sigma)$ with itself
$[U_k''(t)]_{1j}$	$P[\text{BP of } Q_2 \text{ starts at or before } t, Q_1 \text{ non-empty when the 1st cell arrives at } Q_2, X''(t)=k, J(t)=j X''(0)=0, J(0)=1,]$
$\Phi_{c1j}^{1j'}(\tau)$	$E[\text{No of visits of } Q'(\cdot) \text{ to the state } (c1,j) \text{ in } [0,\tau] \mid \text{at time 0 the state was } (1,j')]$
$\Phi_{c1j}^{1j'}(\tau)$	$E[\text{No of visits of } Q(\cdot) \text{ to the state } (c1,j) \text{ in } [0,\tau] \mid \text{at time 0 the state was } (1,j')]$
$W''(\sigma)$	$MN \times 1$ vector whose ℓ^{th} element, denoted as $W_\ell''(\sigma)$, gives the probability that $v''(t) \leq \sigma$ and MMPP 2 is in phase ℓ
$\tilde{W}''(s)$	the LST of $W''(\sigma)$
$W''(0)$	the probability that $v''(t)$ is zero as $t \rightarrow \infty$
$W^{(n)}$	the n^{th} moment of $\tilde{W}''(s)$ evaluated at $s=0$
θ	the invariant probability vector of Q^*
θ''	the invariant probability vector of Q^{**}
$\tilde{U}_{c1}'' , \tilde{U}_1$	the LST of $U_{c1}''(t)$ and $U_1(t)$ evaluated at $s=0$
W_H, W_L	The average queueing delay at Q_2, Q_1
W_T	the average queueing delay of a cell arriving at Q

$\underline{G}^{(N,k)}(t)$	MN×MN matrices whose $(i,j)^{th}$ elements denote P[BP of Q2 is of duration at most t sec and ends with MMPP $\underline{2}$ in phase j given that the BP of Q2 with a capacity of N started with k cells and with MMPP $\underline{2}$ in phase i]
$\underline{G}^{(N,k)}(\ell, t)$	MN×MN matrices whose $(i,j)^{th}$ elements denote P[BP of Q2 is of duration at most t sec, consists of ℓ services and ends with MMPP $\underline{2}$ in phase j given that the BP of Q2 of capacity N started with k cells and with MMPP $\underline{2}$ in phase i]
$\underline{G}_{\ell}^{(n,k)}$	MN×MN matrices whose $(i,j)^{th}$ elements denote P[BP of Q2 of capacity n is of length ℓD sec and ends with MMPP $\underline{2}$ in phase j it started with k customers in phase i]
$\hat{\underline{G}}_{\ell}^{(n,k)}$	N×N matrices whose $(i,j)^{th}$ elements denote P[BP of Q2 of capacity n is of length ℓD sec and ends with MMPP $\underline{2}$ in phase j it started with k customers in phase i]
$\mathfrak{G}^{(n,k)}$	MN×MN matrices whose $(i,j)^{th}$ elements gives the length of the the BP of Q2 with capacity n which ends when MMPP $\underline{2}$ is in phase j given that it started with k cells in Q2 and with MMPP $\underline{2}$ in phase i
X	matrix arrays $X(i)$ whose j^{th} element denotes P[BP of Q2 of capacity n starting with i customers is jD sec]
$\overline{\mu}^n$	A vector whose i^{th} element gives the average number of customers served during a busy period of Q2 of capacity n which starts with the MMPP $\underline{2}$ in phase i
P_{S1}	P[server busy with the customers from $Q1$]

CHAPTER 1

INTRODUCTION

1.1. INTRODUCTION

Broadband Integrated Services Digital Networks (B-ISDN) are envisaged to be the architecture of global networking for the future, integrating diverse services such as voice, video and data. The Public Switched Telephone Networks (PSTN), the Data networks and Narrow band ISDN (N-ISDN) that are in operation today, are not completely satisfactory for multimedia applications and the B-ISDN is the step towards overcoming their limitations. We give a brief account of the evolution of these networks first. Parallel to the evolution of the B-ISDN, the teletraffic theory used to address some of the issues associated with the design and operation of these networks has also undergone a similar evolution. We present a brief look at this next. In this thesis the teletraffic method is used to study a problem which has potential application in B-ISDN. A brief statement of this problem as well as the organisation of the thesis is presented finally.

1.2 THE PRESENT PSTN AND DATA NETWORKS

The Public Switched Telephone network (PSTN) has the largest connectivity over the globe and is more than a century old. Over these years, it has evolved from the electromechanical systems to the present circuit switched digital stored program control exchanges. The initial design goal of PSTN was to provide switched voice communication among its users. Upon recognition of the advantages achieved through the use of digital representation of speech for transport and switching of these signals in digital form, the trunk network of PSTN has evolved into digital transport and switching technology. Parallel to

the evolution of the trunk network of PSTN, the user equipments have also been increasingly digitized with the development of more powerful microprocessors and signal processing techniques. With the proliferation of digital technology, the preferred means of communicating user information, including speech, has become digital. The Integrated services Digital Networks (ISDN), the entirely digital version of the PSTN including the subscriber loop, holds great potential for communicating user information in digital form.

Compared to the PSTN, the data networks are of recent origin. For the transfer of computer data between any two remote users, use of the PSTN on the one hand and the development of dedicated public data networks on the other hand were chosen. The data networks entered for the first time in 1960's under the guise of time shared computer system. The computer vendors in order to increase the market for their costly computing system and at the same time offer the advantages of "economy of scales", set up data networks to access the centrally placed big computers over terminals dispersed over a wide geographical area.

The 1970's brought a tremendous change in the computer networking philosophy (see for example Schwartz [1]). Thanks to the development and advancement of VLSI technology, faster, less expensive processors (CPUs), faster less expensive primary memory and faster, larger, less expensive bulk storage became a distinct reality. The personal computers and work stations became more powerful as well as cost effective and gave the required impetus for the design and implementation of the Local Area Networks (LANs). Interconnection of LANs over a wide area and transport of the data efficiently at high speeds is now increasingly implemented using frame relay and Switched Multimegabit Data Service (SMDS) [2].

1.3 NARROWBAND AND BROADBAND ISDN

The increasing demand for data services and the potential advantages in integrating the switching and multiplexing of voice and data over the PSTN paved the way for the development of ISDN. The computer data may be transported either using the LANs or the ISDN. The latest high speed LANs offer higher user throughputs by two orders of magnitude over the ISDN. Hence, the promise of ISDN is not in providing the highest possible speed or setting up a performance record in some other way. Instead, the universal connectivity with sophisticated signalling and control that ISDN offers, makes it the well defined and natural evolutionary path of PSTN [3].

In the 1980's, ISDN became operational in some parts of the globe. The essential functional elements of public ISDN [4] are shown in Fig 1.1. These are

- 1 ISDN local exchanges with digital subscriber lines
- 2 Common channel signalling capabilities (CCITT signalling system No 7) for transferring signalling information between exchanges
- 3 Physical connections with circuit switching
- 4 Virtual connections with packet switching
- 5 Specialized equipment for additional functions

Two interfaces are defined for ISDNs: the Basic Rate interface (BRI) and Primary Rate interface (PRI). The BRI has 2 B channels of rate 64 kbps and 1 D channel of rate 16 kbps. The PRI has either 23 or 30 B channels and a 64 kbps D channel. The BRI and PRI ISDN offerings are often collectively referred to as narrowband ISDN (N-ISDN) to distinguish them from B-ISDN. The data rates associated with N-ISDN are inadequate for many applications of interest [5]. On the one hand, the BRI providing 64 kbps is not a large improvement over the modem rates of 9.6 and 19.2 kbps widely available. For data transmission, the

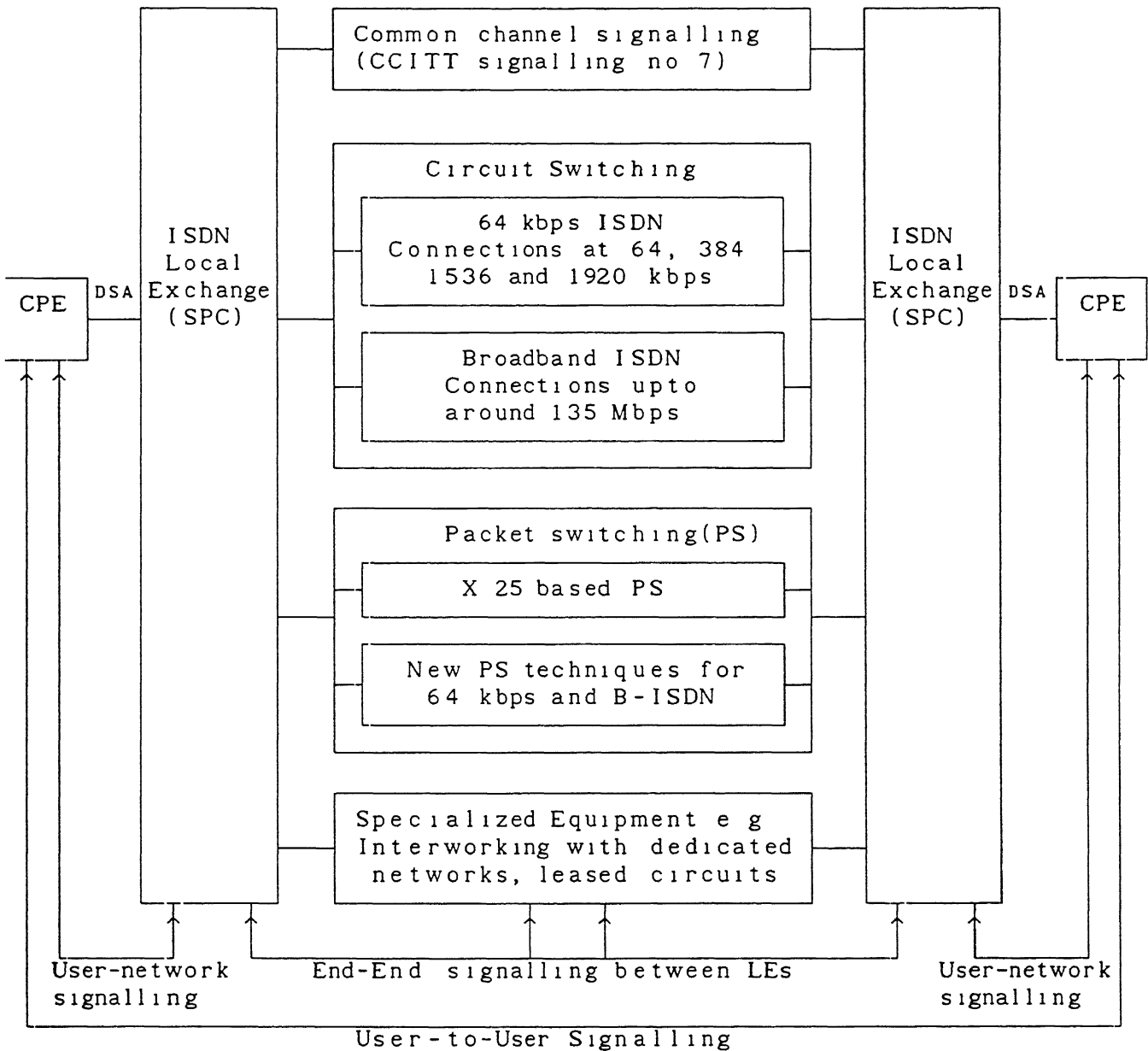


Fig11 Basic functional elements of a public ISDN

CPE- customer premises equipment

DSA- Digital subscriber access

LE- Local Exchange

SPC- Stored Programme Control

BRI rates are inadequate and much less than the 10Mbps and more available from LANs. The PRI rates on the other hand are no superior than the point to point T1 offerings in use today. For applications such as interconnection of LANs, video, and image, these interfaces are found to be inadequate and this led to the need for the design and development of Broadband ISDN (B-ISDN) capable of providing flexible customer bandwidths upto hundreds of Mbps.

It is now acknowledged that the technology of choice for B-ISDN multiplexing and switching is cell relay or more commonly referred to as the Asynchronous Transfer mode (ATM). Initially, the cell based packet switching, later called ATM, was driven by the research and development of high speed switch fabrics. To support the packet switching function at very high speed, hardware solutions were mandatory. The fixed size data units were easier to handle than the variable size packets. In addition to this, ATM is a universal, service independent switching and multiplexing technique that can utilize any of the underlying digital transmission technologies and transmission speeds.

The B-ISDN is expected to cater to a variety of application such as conversational services, retrieval services, messaging services and distribution services. The Quality of service (QoS) requirements of some of these applications (see for example [6]) are quite different as shown in Fig 1.2. The physical layer for the B-ISDN is the Synchronous Optical Fibre Network (SONET). SONET has laid out a hierarchy of transmission speeds from 51.8 Mbps upto 13.27 Gbps and above. These very large communication bandwidths have caused a wealth of research and experimentation to take place in fast packet switching.

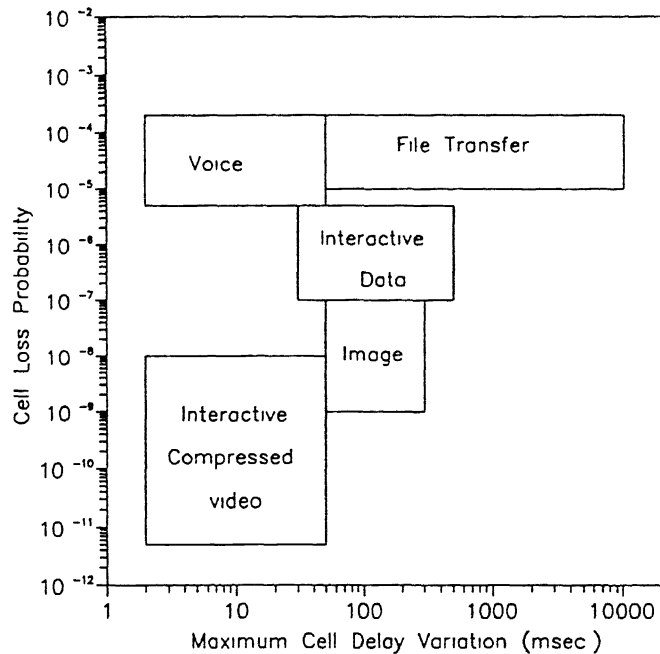


Fig 1.2 Approximate ATM traffic performance requirements

1.4 TELETRAFFIC ENGINEERING

We have so far considered the evolution of the communication networks in brief. It is clear that the twin objectives of providing cost effective service and offering the most satisfactory performance to attract more and more customers are very much in the objectives of communication networks. In a communication network, a major decision regarding how much of resources should be made available to ensure a particular level of customer satisfaction and cost effectiveness has to be taken after considering several complex and at times conflicting factors. For example, in a PSTN, for a call to get accepted, switching elements, digit receivers, interstage switching links, call processors and trunks between the exchanges should all be free. The load or the traffic pattern varies over the day with heavy traffic at certain times and light load at some other time. The traffic engineering also called as the teletraffic theory provides the basis for the analysis required for arriving at these decisions and design the network. It enables one to determine the ability of a telecommunication network to carry a given traffic at a particu-

lar call loss probability. It provides the means to determine the quantum of equipment required to provide a particular Quality of Service (QOS) for a given traffic pattern and volume.

Several books have been written on traffic engineering (see for example Syski [7], Bear [8]). In traffic engineering, the statistical description of the traffics arriving at the system and the holding time of the system by the individual customers are modelled and the performance of the system is analyzed using methods like queueing analysis. With currently available powerful computers, it has become possible to use more complex and realistic models to evaluate the system performance and switching architectures. With the advancement in the VLSI technology, signal processing chips capable of achieving bandwidth reduction of 20:1 have become a reality. With these chips, the subband codings have been used to "squeeze in" the maximum information out of the available bandwidth. This makes the input stream to the system to be often too complex that they are not always amenable for being modelled using simple input processes. This again warrants more complex models for the analysis.

The computational powers have increased so much that the queueing analysis is not restricted to the study of the performance of the networks off-line but is also used for on-line evaluation. For example, for admitting a new call in an ATM network, methods based on queueing analysis is also proposed as a means for verifying whether the new call can be accepted without degrading the overall quality of service of the ongoing calls. This requires the development of simple as well as accurate models to capture the characteristics of the complex arrival streams. The Poisson models which are mathematically tractable and at the same time reasonably accurate for the analog telephone networks are not completely satisfactory for modelling the bursty traffic

arriving at the ATM networks. More complex arrival processes like Markov Modulated Poisson Process are considered for the study of these networks.

The availability of the best computing facility has also altered the directions of some of the researches in teletraffic engineering. In 1970's, alternative techniques for solving the teletraffic problems were proposed and are referred to as Matrix Geometric - Matrix analytic methods (see for example Neuts [9]). These techniques obtain the solution of some of the complex problems in a form that is best suited for numerical implementation.

1.5 STATEMENT OF THE PROBLEM

In view of the need for the study of queues with more complex arrival streams and the desirable features of the matrix analytic approach, the study of a non-preemptive priority system in which the arrivals are modelled as Markov Modulated Poisson Process using the matrix analytic method has been considered in this thesis. Under a set of assumptions, the evaluation of the distribution of queue lengths, moments of queue lengths, the average queueing delays for the customers of different priority classes, the percentile of queueing delays of the highest priority queue is considered for the cases where capacity of the queues are either infinite or finite. The results obtained in this thesis have potential application in Asynchronous Transfer Mode (ATM) networks.

1.6 ORGANIZATION OF THE THESIS

The organization of the rest of the thesis is as follows.

In chapter 2, a review of the previous work and the motivation for undertaking the work carried out here are presented. In section 2.2, in order to motivate the need for the study of a priority system with more complex arrival

streams an introduction to ATM networks and the role played by the queueing analysis in the design operation and maintenance of these networks is highlighted. The need for the incorporation of service priority in the queueing model is examined. In section 2.3, a review of the source models used for the queueing analysis of modern communication networks is given. In section 2.4, a review of the previous work carried out in the study of queueing systems with priority is given. In section 2.5, an overview of the method adopted in this thesis for the study of a non-preemptive MMPP/D/1 priority system with dual priority classes and infinite or finite waiting space is presented.

Chapter 3 considers the choice of appropriate embedded points for the evaluation of the queue length densities at the low and high priority queues. A study of some of the characteristics of the transition probability matrices pertaining to these queues is presented. In section 3.1, the assumptions used for the prioritized queueing system are given. A simplified notation for the specification of the semi Markov chains pertaining to the low and high priority queues (denoted as Q_1 and Q_2) is presented in section 3.2. Section 3.3 describes the structure of the state transition probability matrices and gives the defining equations for obtaining the elements of these matrices. In section 3.4 and 3.5 equations for determining the elements of the state transition probability matrices pertaining to Q_2 and Q_1 respectively are obtained. Section 3.6 considers the evaluation of the transform of these elements.

In chapter 4, some characteristics of the busy periods of Q_1 and Q_2 are studied. In section 4.2, the so-called $G'()$ and $G''()$ matrices whose role is analogous to that of busy period distribution in an M/G/1 queue is introduced and some of their properties are considered. In section 4.3, a recursive procedure for the evaluation of the busy period distribution of Q_1 and Q_2 are presented. In section 4.4, application of this recursive procedure for the

computation of the matrices $P'()$ and $P''()$ whose elements denote the probability of arrivals of customers at Q1 and Q2 from the MMPPs and their phase transitions are considered. In section 4.5, computation of the average length of the Busy period (BP) and the average number of customers served in a BP of Q1 and Q2 are considered. Section 4.6 considers the evaluation of the average lengths of busy cycles of Q1 and Q2.

In chapter 5, some details on the evaluation of the queue length density (QLD) of Q1 and Q2 and the numerical results are given. Section 5.2 considers the computation of x'_0 and x''_0 which denote respectively (the joint probability of Q1 being empty and MMPP to Q1 being in a particular phase at a Q1 departure instant) and (the joint probability of Q2 being empty and MMPP to Q2 being in a particular phase at a Q2 departure instant). Section 5.3 considers the computation of x'_1 and x''_1 which denote respectively (the joint probability of Q1 having one customer in the system and MMPP to Q1 being in a particular phase at a Q1 departure instant) and (the joint probability of Q2 having one customer in the system and MMPP to Q2 being in a particular phase at a Q2 departure instant). Section 5.4 considers the computation of the probability of Q1, Q2 being empty at an arbitrary time instant. In section 5.5 the evaluation of the moments of the QLDs are considered. Some details on the evaluation of the QLDs using (i) Gaussian elimination (ii) Toeplitz matrix inversion (iii) recursive procedure, are considered in section 5.6. It also considers some simple extensions for computing (i) QLDs of Q1 and Q2 when Q1 has a finite size (ii) QLD of Q2 alone without evaluating the QLD of Q1. In section 5.7 approximate models for the computation of QLDs of Q1 and Q2 are considered. In section 5.8 the numerical approximations made for the computation are discussed. In section 5.9, some details of the simulation routine for the validation of the computational results are presented. In section 5.10 the

results obtained through computation and simulation are presented

In chapter 6, the evaluation of the LST of the virtual waiting time at Q2, the average delays and average queue lengths at Q1 and Q2 are considered. In section 6.2, expression for the cumulative density function of the virtual waiting time at Q2 is obtained. In section 6.3, evaluation of the LST of the expression obtained in section 6.2 is considered. In section 6.4, extension of the results of section 6.2 and section 6.3 for the approximate model proposed in section 5.7 is discussed. In section 6.5, some details on the computation of the average queueing delays and the approximations made are discussed. Computation of the average queue lengths at Q1 and Q2 are also considered. Section 6.6 discusses the extension of these results for the M/D/1 non-preemptive priority system. In section 6.7 the results obtained through the computation is compared with the simulation results.

In chapter 7, the study of the non-preemptive priority system in which the high priority queue has a finite size is considered. In section 7.2, the embedded semi-Markov sequence and the state transition matrices pertaining to the low priority (Q1) and high priority (Q2) queues for the case where Q2 has a finite capacity is considered. Section 7.3 discusses some of the properties of the busy period. A recursive procedure for the computation of the busy period distribution of the finite capacity Q2 as well as a technique for minimizing the time for computing the busy period distribution is discussed. Computation of the average number of customers served during the busy period of Q2 is also considered. In section 7.4, the computation of the QLDs at Q1 and Q2 are considered. As in chapter 5, an approximate model is proposed for the finite capacity system. The numerical results obtained through the computation are compared with those obtained using simulation.

Chapter 8, studies a non-preemptive MMPP/D/1 priority system with more than

two priority classes. In section 8.2, evaluation of the queue length density for the case where there are three priority classes and three dedicated buffers is considered. Extension of these results for the case where the number of priority classes are more than three is discussed. In section 8.3 the evaluation of the queueing delay as well as the average queue length for the case of three priority classes is considered first. The extension for four or more priority classes is indicated.

Chapter 9 summarizes the results obtained and the conclusions drawn.

REFERENCES.

- 1 M. Schwartz, "Telecommunication Networks", Addison-Wesley, USA, 1988
- 2 M. Irfan Ali, "Frame relay in Public Networks", IEEE Commn Magazine, March 1992, pp 72-78
- 3 Nuri Dagdeviren, J. A. Newell, L. A. Spindel and M. J. Stefanick, "Global Networking with ISDN", IEEE Communication Magazine, June 1994, pp 26-32
- 4 Peter Bocker, "ISDN - The Integrated Services Digital Network", Springer Verlag, Berlin, 1988
- 5 L. Kleinrock, "ISDN - The path to Broadband Networks", Proc IEEE, Vol 29, No 2, February 1991, pp 112-117
- 6 D. Hong and T. Suda, "Congestion Control and Prevention in ATM Networks", IEEE Network Magazine, July 1991, pp 10-15
- 7 R. Syski, "Introduction to congestion theory in Telephone systems", Oliver and Boyd, London, 1976
- 8 D. Bear, "Principles of Telecommunication Traffic Engineering", Peter Peregrinus Ltd, London, 1976
- 9 M. F. Neuts, "Matrix-Geometric Solutions in Stochastic models. An algorithmic approach", Johns Hopkins Univ Press, Baltimore, 1981

CHAPTER 2

REVIEW OF THE PREVIOUS WORK AND AN OVERVIEW OF THE PROBLEM

2.1 INTRODUCTION

Integration of various diverse services like voice, video and data on a single network can be achieved in a Broadband Integrated Services Digital Network (B-ISDN). In recent times, this type of integration [1] has been felt desirable in view of the wider connectivity and economies of scale that can be achieved with this approach. A major difficulty in such an integration is that the system should take into account the fact that these services will generally have widely different quality of service (QOS) requirements. For example, traffic such as those generated from voice and some video applications can tolerate some loss of packets but are delay sensitive. Data traffic generated by computers is loss sensitive but can usually tolerate large, random delays. Some sub-band coded video services are both loss and delay sensitive. In order to integrate these services and at the same time meet their conflicting QOS requirements, incorporation of some kind of priority in the underlying communication network may become necessary.

Various schemes have been proposed in the literature to integrate services such as voice and data, with different QOS requirements, on the same network. This problem has been typically considered for multiple access networks such as Local Area Networks (LANs) and Radio Networks [2-7]. Most of these techniques use some method for suitably incorporating priority to achieve the different QOS requirements for voice and data services. However, the design issues involved in incorporating priorities in a multiple access network and in B-ISDN are quite different. In multiple access networks, the design problem becomes one of designing suitable protocols to discriminate between the diffe-

rent services in awarding permission to access the shared communication medium, this is generally done depending upon their delay sensitivity requirements. In B-ISDN, the problem becomes one of designing suitable queueing disciplines for the different services and studying the extent to which the QOS of the various services are satisfied by these queueing disciplines.

In our present work, we are interested in the study of a prioritized queueing system which has a potential for use in ATM networks. In order to motivate the study of the non-preemptive MMPP/D/1 queueing model presented in our present work and to appreciate its relevance in the ATM context, we give next a brief review of some of the principles, characteristics and issues that arise in ATM.

2.2. INTRODUCTION TO ATM NETWORKS

2.2.1. PRINCIPLES OF ATM

Asynchronous Transfer Mode (ATM) networks are high speed networks designed to integrate a wide variety of different communication services (eg services like voice, video and data) using small fixed size packets (53 bytes) and a statistical Time Division Multiplexing technique [8-10], these small fixed size packets are referred to as cells in ATM. The statistical nature of ATM and the absence of fixed reservations sets it apart from the older Synchronous TDM (STM) approach. In the STM scheme, when a call is set up between two parties, the information from user1 to user2 (and vice versa) always appears in a fixed TDM slot with respect to the sync bit during the entire holding time of the call. This has been shown in Fig 2.1. In this figure F denotes the slot used for indicating the beginning of the frame. With respect to this slot

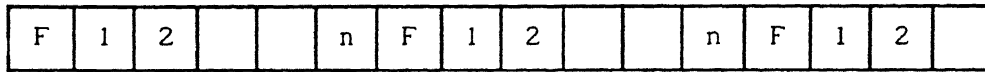


Fig 21 Synchronous Time division multiplexing

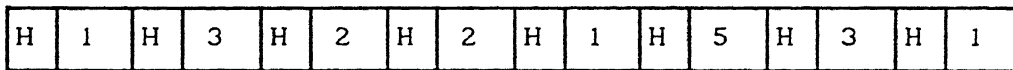


Fig 22 Statistical Time division multiplexing

the information corresponding to different channels appear at the same slot in each frame. Therefore there is no need to transmit the destination address of each data packet/cell while the call is in progress. However such fixed allocation of slots/channel is inefficient as the channel is also forced to be idle during pauses or idle periods of the source involved in the call.

ATM achieves better utilization of the channel by allocating the slots to the connections only when there is activity at the source and the cells are actually available for transmission. This has been shown in Fig 22. In this figure H denotes the header of a cell from the 1th call. It can be noted that the cells from the different calls are not transmitted in the same sequence. Since a large number of sources are multiplexed in ATM, an incoming cell from a particular connection may find on arrival that the link is currently transmitting the cell corresponding to some other connection, this cell is then put into the ATM MUX buffer for later transmission. The cells are normally transmitted on a FCFS basis. Note that even if the cells originate periodically from the source, they may spend a random time in the nodal buffer and may subsequently appear to be aperiodic at the destination. These fluctuations need to be smoothened out by "storing and playing". Every cell must also carry information regarding its destination to enable it to be routed properly in the system.

2.2.2 SPECIFIC CHARACTERISTICS OF ATM NETWORKS

The Synchronous Optical Fibre Network (SONET) has been chosen as a physical layer media for ATM networks. For ATM/SONET networks, bit error rates are guaranteed to be less than 10^{-8} . Therefore, error recovery on a link-by-link basis is not required in these systems. In these systems, bit transmission rates of the order of 150 Mbps is typical and hence the queueing delays in the nodes become negligible compared to the propagation delays between the nodes. This has been shown in Tables 2.1 and 2.2. Table 2.3 shows that buffers of moderate size are enough to ensure very small buffer overflow probabilities in these systems. These results are obtained in [11] using M/D/1 model for the queueing system.

ρ	56Kbps	1.54Mbps	150Mbps
0.1	410	15	0.1
0.3	1620	59	0.6
0.5	3790	137	1.4
0.7	8830	321	3.2
0.8	15140	550	5.6
0.9	34070	1238	12.7
0.95	71920	2613	26.8
0.99	373700	13612	139.6
transmission time	7570	275	2.83

packet size 53 bytes

Table 2.1 Queueing delay and Transmission time (in $\mu\text{sec.}$)

Distance[miles]	Propagation delay [μs]
30	200
300	2000
3000	20000

Table 2.2 Propagation Delay

ρ	32	64	128
0 5	$< 10^{-13}$	$< 10^{-13}$	$< 10^{-13}$
0 6	$< 10^{-13}$	$< 10^{-13}$	$< 10^{-13}$
0 7	$0 71 \times 10^{-10}$	$< 10^{-13}$	$< 10^{-13}$
0 8	$0 14 \times 10^{-6}$	$0 15 \times 10^{-12}$	$< 10^{-13}$
0 85	$0 46 \times 10^{-5}$	$0 18 \times 10^{-9}$	$< 10^{-13}$
0 9	$0 11 \times 10^{-3}$	$0 15 \times 10^{-6}$	$0 26 \times 10^{-12}$
0 95	$0 18 \times 10^{-2}$	$0 68 \times 10^{-4}$	$0 10 \times 10^{-6}$
0 99	$0 11 \times 10^{-1}$	$0 37 \times 10^{-2}$	$0 81 \times 10^{-3}$

Table 23 Packet loss probability as a function of buffer size

2 2 3 ATM CELL STRUCTURE

The ATM cell format is shown in Fig 2 3 Each ATM cell has a 5-byte header and a 48-byte information field

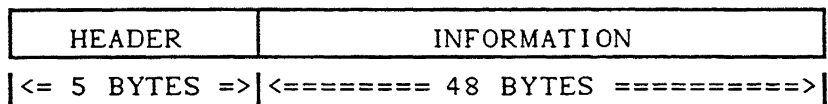


Fig.2 3 Organization of the cell

The various fields of the header and the number of bits allotted for each field are as follows

1	Generic flow control (GFC)	4 bits
2	Virtual path (VP) field	8 bits
3	Virtual channel (VC)	16 bits
4	Pay load type field	2 bits
5	Loss Priority	1 bit
6	Reserved	1 bit
7	Forward error correcting	
	header check sequence	8 bits

The header of a cell at the network node interface does not have a gfc field, instead it has a longer VP field of 12 bits

2.2.4 LAYERED ARCHITECTURE FOR ATM NETWORKS

An ATM network has a four layered architecture as shown in Fig 2.4

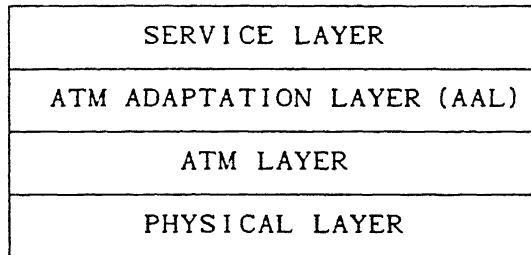


Fig 2.4 Layered architecture for ATM networks

As mentioned earlier, the physical layer is usually based on SONET with transmission rates of the order of 150 Mbps. The ATM layer consists of the ATM multiplexer, buffer and the ATM switch. The function performed by this layer is kept minimal so that these functions can be done quickly and effectively using hardware support. The adaptation layer attends to the specific needs of connection-oriented as well as connection-less services, end-to-end synchronization, flow control, error recovery, information segmentation and reassembly, and supervision of the Quality of Service (QoS) offered to the various services. The service layer depends on the specific service for which a particular connection is used. Typical services include constant bit rate voice/video, variable bit rate video, compressed voice and digital data.

2.3 SOME OF THE DESIGN ISSUES IN ATM NETWORKS

In this section we shall consider some design issues associated with the management of traffic in ATM networks. For brevity, we shall not consider here the other design issues like Switching, Routing etc., these issues are discussed in detail in [8],[11-16].

2.3.1 CALL ADMISSION CONTROL

As mentioned earlier, instead of allocating the bandwidth permanently for each call, it is allocated in a dynamic manner in ATM. Taking into account the bursty nature of cell arrivals from each call in an ATM network, more calls are allowed to co-exist than is actually permissible if all calls were to be allocated fixed bandwidth. The cells arriving at the ATM multiplexer are queued in a buffer and served usually on a FCFS basis as shown in Fig 2.5.

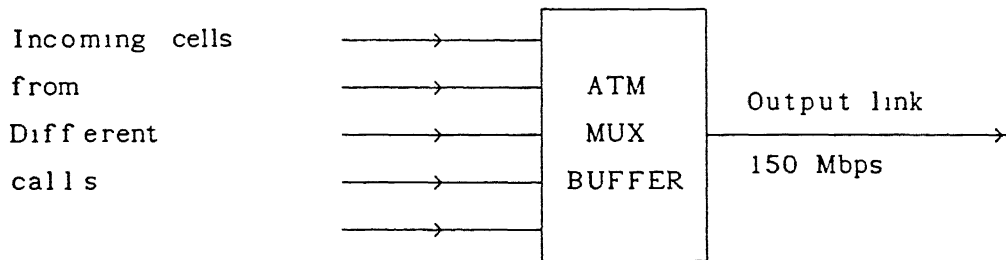


Fig 2.5 ATM multiplexer and the buffer

If a large number of calls are allowed to coexist, then the queueing delay in the ATM buffer becomes large and overall end-to-end delay may reach objectionable limits. Even if delays are not very critical, the fact that buffer sizes are finite will cause problems when the number of simultaneous calls are large. In this case, the entire buffer may become full and an incoming cell will then have to be discarded. Hence the maximum permissible queueing delay and the acceptable cell loss probability restrict the maximum number of calls that can be simultaneously allowed at the ATM mux/buffer without degrading the QoS (cell loss probability, queueing delay) requirements of the individual calls. Call Admission Control procedures (CAC) are developed to handle this problem[17]. At the time a new call is requested, the QoS requirements of the call as well as its traffic parameters (peak cell rate,

average cell rate, peak duration etc) are furnished to the ATM mux. The CAC procedures are invoked to check whether there is adequate bandwidth to support the new call satisfactorily without degrading the QOS of the other on-going calls. Since the CAC is invoked for every incoming call, it must be both accurate and simple enough to be executed on-line. The CAC procedures can either use queueing analysis or use virtual bandwidth concepts to ascertain whether an incoming call can be accepted. In the former case, assuming some suitable models for the traffic sources undergoing service and the incoming traffic, queueing analysis is carried out to see whether the average queueing delay and the cell loss probability that results using the finite buffer is below the QOS requirements of the calls [18-23]. As this has to be done in real time, only simple traffic models can be used for the analysis. In the other method, the traffic parameters of the calls are converted into some equivalent bandwidth requirements and then the problem becomes one of assessing whether the total bandwidth requirement is below the bandwidth available on the outgoing link [24-27].

In ATM networks, cell loss requirements are very stringent and cell loss probability due to buffer saturation (overflow) should not exceed 10^{-9} . Studies [28] have been carried out to see which of the traffic parameters need to be considered for CAC for various buffer sizes. It has been found that only the peak and mean rates need to be considered for CAC when the buffer size is small, when the buffer size is large, CAC procedures have to consider the burst length as well as the probability distribution of the burst, in addition to the peak and average cell rates of the calls.

Two methods of CAC have been considered in [29]. In the virtual channel method, each call for a particular source-destination pair has to get its call accepted at each of the intermediate nodes invoking each of their own CACs.

alongwith their on-going call statistics. The call is accepted only when all the nodes have the required bandwidth. In the virtual path (VP) method, the bandwidth is reserved for various destinations in an apriori fashion at each node. A new incoming call request needs to be processed only at the source node and the CAC need not be invoked in the intermediate nodes. The bandwidth allocated apriori to each of the destination nodes is done in anticipation of the traffic for them and has to be updated from time to time depending on the actual utilization statistics. An incoming call which does not get the required bandwidth will be denied access. It is claimed that the VP method reduces the control costs by 30%. However, the bandwidth allocated to one virtual path may be under utilized even while calls made through other virtual paths are being rejected for lack of sufficient bandwidth. This may offset the advantages of statistical multiplexing.

2.3.2 TRAFFIC PARAMETER CONTROL (POLICING FUNCTION)

CAC will be effective in ensuring the QOS of each call only if the sources restrict the traffic within the negotiated range of parameters. In practice, the actual values of the traffic parameters (peak cell rate, peak duration, average cell rate) from a source may be different from what was negotiated, this may be due to (a) inadequate apriori statistics about the traffic, (b) malfunctioning of equipment or (c) by deliberate attempts to understate the parameters to avoid higher tariff rates. In order to ensure the QOS for the disciplined user, a policing function is introduced to monitor the traffic from each source and mark the cells whenever any parameter range is violated for that source. Such marked cells may be transmitted to the destination only if there is no congestion in the intermediate nodes. Otherwise, the marked cells are dropped. Additional revenue may also be charged at penalty rates for the marked cells that actually reach the destination.

A popularly implemented policing scheme, known as the leaky bucket mechanism (LB), consists of a counter which is incremented by 1 each time a new cell comes and is decremented periodically as long as the counter value is positive [30]. If the momentary arrival rate exceeds the decrementation rate, the counter value starts increasing. It is assumed that the source has exceeded the admissible parameter range if the counter value reaches a predefined limit. At that time, all subsequent cells are discarded or marked (as appropriate), until the counter falls below its limit once again. Single LB mechanism is inadequate to simultaneously detect both peak rate and average rate violations, two stages are employed to detect both these violations, one after another. The LB mechanism is not effective in detecting burst duration violations. Methods other than the Leaky Bucket mechanism have also been considered [8]. These are however more difficult to implement.

2.3.3 CONGESTION CONTROL IN ATM NETWORKS

As noted in section 2.3.1, if a large number of sources become active simultaneously, the nodal (ATM mux) buffers saturate and network congestion results. In view of the large bit rates used, the transmission time of a cell is much smaller than the propagation delay between the source and its destination. Therefore, by the time the first cell reaches the destination, the source would have already transmitted thousands of cells. Because of this, efforts of the destination to reduce the input rate would be ineffective in controlling the cell arrival rate. Moreover, since high bit rates are used, the protocols at the ATM nodes should be simple enough to be implemented in hardware. Hence, congestion control mechanisms should be implemented on an end-to-end basis rather than on a link-to-link basis [31].

Preventive control is exercised in ATM networks to protect the network from reaching unacceptable levels of congestion. The call admission control and the traffic parameter control discussed earlier fall in this category. As mentioned earlier, when congestion occurs, the marked cells will be dropped to reduce congestion.

In addition, the sources themselves can organize their cells into ones which are important and those which can be dropped in case of network congestion (without too much degradation in the overall performance of the associated service). As an example, consider the following.

Consider voice sources where we assume that the Adaptation layer header is 4 bytes long, and that the actual voice information is 44 bytes/cell. The time taken for 44 bytes to be available for transmission from a voice source coded using ADPCM is 11 msec. Instead of generating a cell every 11 msec, samples corresponding to 22 msec will be organized into two cells with all the MSBs of the samples in one cell and LSBs in the other cell [32]. The first cell is identified as non discardable and the other as discardable by appropriately setting the loss priority field in the ATM header. An alternative to the above scheme is to put all the odd speech samples in one cell and the even samples in the other cell and mark one of them as droppable.

Three methods are considered for network wide congestion control [33]. As discussed earlier, in the first method, a capability for selectively shedding the cells under congestion conditions is achieved by marking the cell loss priority (CLP) indicator in the cell header. This bit will be made 1 either by the source itself or by the policing device. The cells with CLP=1 will be discarded in case of congestion.

The congestion condition existing along a particular virtual circuit/path may also be conveyed to the destination through the Forward congestion

indicator (FCI) in the header. This bit will be set to 1 if any intermediate node is congested. The combination of the FCI and CLP indicators provides information on the status of congestion in the network. At low levels of congestion, cells with CLP=1 and FCI=1 will be received at the destination. At higher levels, cells with CLP=0 will be received with FCI=1 and cells with CLP=1 will probably be lost.

The third technique, is to generate backward notification to the originating nodes on the congestion encountered for cells with CLP=0. However, this technique requires the network to originate congestion messages and hence increases the overload traffic.

2.3.5 INCORPORATION OF PRIORITY IN ATM NETWORKS

SPACE PRIORITY MECHANISMS

Consider a situation where calls with different QOS requirements are being served at the ATM mux. If all calls were to be treated alike, then the maximum number of calls that can be allowed to coexist will be based on the most demanding QOS specification. For example, signalling and sub-band coded video involve vital cells which must be received by the adaptation layer. On the other hand, voice and data communications, which represent the majority of calls, could cope with higher cell loss rates and delay jitters. An efficient way to accommodate maximum amount of traffic is to have two bearer services - one for the vital, highly loss sensitive cells and the other for the ordinary cells. Incorporation of such a technique is known as the space priority mechanism. Space priority can be incorporated either at the call level or at the cell level.

Two popular schemes for implementing space priority in ATM networks are partial buffer sharing and the push out mechanism. In both of these, the loss

sensitive as well as ordinary cells arrive at a single buffer. In partial buffer sharing, the incoming ordinary cells are stored in the buffer only if the number of cells in the buffer is below a particular threshold. However, the loss sensitive cells will be accepted as long as the buffer is not full. While this scheme is easy to implement, a higher priority cell may be lost because it arrives late compared to ordinary cells already in the buffer. Study of the partial buffer scheme has been carried out under various assumptions about the input arrival statistics in [34-36].

In the push out scheme, an incoming cell will be accepted into the buffer if there is space. The ordinary cells which arrive at the filled buffer will be discarded. However, the loss sensitive cells will still be accommodated in a full buffer as long as space for it can be made by pushing out any lower priority ones which may already be in the buffer. The incoming loss sensitive cell will be discarded only if the buffer is full and there are no lower priority cells in the buffer. Compared to partial buffer sharing, this is more effective in ensuring a low cell loss probability for the higher priority cells. However, the implementation complexity is more for this scheme. Evaluation of the cell loss probability for both the priority classes is considered in [37-39].

INCORPORATION OF TIME PRIORITY

In ATM networks, the need may arise in some cases to offer time priority (i.e. transmission of cells from the buffer on a priority basis instead of using FCFS discipline) either to enhance a space (loss) priority mechanism or to provide certain classes of traffic with faster network transfer rates with smaller transfer delays and lower delay jitters. For example, in manufacturing environments, alarms and real time control informations are time critical and

may be carried at a higher priority. If an ATM switch with an input buffer is used, then it has been found that incorporation of priorities can actually increase the throughput of the switch [38]. In real-life systems, it may be necessary to provide higher priority service to transactions which are associated with network monitoring and control. This may indeed be the main driving force behind incorporating time priority mechanisms in future ATM systems.

2.3.6 TRAFFIC MODELS FOR ATM SOURCES

For the analysis of queueing models, one of the simplest models used for characterizing the arrival process is the Poisson process. This has the advantages of leading to simple queueing analysis and is fairly accurate if the number of sources are large. Moreover, some data sources can indeed be modelled as Poisson sources reasonably well. However, for the bursty traffic arriving at an ATM mux, the variance to mean ratio is greater than unity, and hence the Poisson process model is not accurate [40].

Another model commonly used to characterize a source is the on/off model. In this model, the source is assumed to emit cells periodically in the on state, it does not emit cells in the off state. The on and off durations are generally assumed to be exponentially distributed. The parameters of the on/off source model can be suitably obtained from measurements made on the actual traffic. A single on/off source may be inadequate to represent all types of sources. In a generalized on/off source model, the source output is represented as the superposition of N on/off sources [41]. The parameters of the generalized on/off source model for some typical sources are given in Table 2.4.

SOURCE TYPE	N_S	T_{ON}	AR	P
POISSON	∞	T_c	0	0
CBR	1	*	1	*
VBR VOICE	1	0.35 msec	0.35	64 kbps
VBR CATV	30	33 msec	0.38	44.1 Mbps
VIDEO CONFERENCE	15	33 msec	0.32	1 Mbps

Table 2.4 Parameters of the generalized on/off source model for some typical sources

In this table CBR denotes the constant bit rate sources (uncompressed voice, video), VBR denotes the variable bit rate source and T_c denotes the cell transmission time. Some of the important parameters of the on/off source model are

peak bit rate	P
average on duration	t_{on}
activity ratio(% of on duration)	AR
number of identical sources required	N_s

In Table 2.4 the values of the parameters marked as (*) depends on the particular source.

Another alternative is to model the arrival process as a Markov Modulated Poisson Process (MMPP). MMPP is a doubly stochastic Poisson process. In this model, the arrival process is assumed to be in one of n possible states and in each state i , arrivals are assumed to come from a Poisson process with mean arrival rate λ_i . The state of the arrival process in turn is governed by an n -state Markov process. The limiting case of a 2-state MMPP in which the arrival rate in one state is zero is referred to as an Interrupted Poisson

Process(IPP) This model has been used to model overflow systems in [41-42] The MMPP model can be used for voice as well as video sources Further, superposition of two MMPPs also results in a new MMPP process whose parameters can be obtained from the parameters of the constituent processes A two state MMPP model is adequate for modelling the voice sources as shown in [44,45] Approximate 2 state MMPP model for video sources has been attempted to arrive at practical algorithms for CAC and other control procedures in [18] However, the numerical procedures using the MMPP model has the disadvantage of slow convergence for a bursty traffic at high traffic intensities [46] The computational complexity also grows rapidly with an increase in the size of the state space In spite of this, MMPP has been widely used in the analysis of ATM networks [18-19], [39], [47-51]

There are several approximation techniques proposed for choosing the parameters of the MMPP [45,47,49,50] We shall consider the method proposed in [45] in some detail A 2-state MMPP has 4 parameters which need to be chosen These parameters are obtained by matching the statistical characteristics of the composite arrival process from measurements of

1. The mean cell arrival rate
- 2 The variance to mean ratio of the number of arrivals in a short time interval
3. The long term variance to mean ratio of the number of arrivals
- 4 The third moment of the number of arrival of cells in the short term

As the number of voice sources are increased, λ_1 and λ_2 become closer, in the limit, when the number is very large, they become equal and the model reduces to the Poisson process

Another popular method used to model the composite traffic from ATM sources is the Fluid Flow model Fluid flow models assume the cells to be generated

continuously rather than at individual instants. The system dynamics is specified in terms of first order differential equations as a function of the probability density function of the system state [52,53]. By assuming some appropriate boundary conditions, the probability of the system state can be evaluated. While the burstiness and the correlation in the input stream can be represented by a fluid flow model, it cannot take care of periodicities in the input stream or account for the discrete nature of the arrivals. It also cannot capture the stochastics of the input and tends to be inaccurate for small buffer sizes [49].

In addition to these, several models like the Bernoulli process, Interrupted Bernoulli process, Switched batch Bernoulli process, Discrete Batch Bernoulli process etc. have been proposed in the literature for modelling ATM sources [54-56]. Autoregressive Models (AR) have also been proposed to characterize the arrivals from video sources. However, this model is suitable only for simulations and not for a queueing analysis. As the ATM system is a slotted system, the superposed traffic can also be modelled as a Markov chain with a known transition probability matrix. This approach has been used for dimensioning a buffer in an ATM node in [57,58].

2.4. REVIEW OF THE PREVIOUS WORK

Having reviewed some of the characteristics of the ATM networks and some design issues involved, We present a summary review of the work done to study the behaviour of prioritized queues. However, we restrict ourselves to methods and models which have potential applications in ATM networks and to systems where only service priorities are being considered.

In [59] Niu et al propose a partial preemptive priority (PPP) for the call level and a selective packet discarding strategy for the cell level to handle

the delay sensitive and non-delay sensitive cells in ATM networks. Two priority levels are assumed both for the call level and for the cell level. The call level priority is used to allocate the Cell Assembler, Disassemblers (CLADs) for the individual calls. For the call level, an $M_1, M_2 / M_1, M_2 / s(\infty, s)$ model is assumed where M_1 and M_2 refer to the arrival processes modelled as Poisson processes corresponding to the low and high priority calls respectively. The storage space for the low and high priority calls are assumed to be ∞ and s respectively. The high priority calls that arrive when there are already s high priority calls in progress are blocked. Using the matrix geometric method, the probability of blocking for the high priority calls and mean waiting time of the low priority calls are obtained. For the cell level, an $M, MMPP^{[x]} / E_k / (\infty, N)$ queueing model is used. For the loss insensitive cells, a batch MMPP with size=2 is assumed. Each batch consists of 2 cells with one corresponding to the most significant (MSP) and the other corresponding to the least significant part (LSP) of the original information. The buffer size for the loss sensitive traffic and non loss sensitive traffic are assumed to be infinite and finite respectively. When the buffer size is below m_1 , both the cells of a batch will be accepted. Above m_1 , only the MSP cell of the batch will be retained. An Erlangian distribution with 30 phases or more is assumed for the server in order to approximate the constant service time/cell. Using this model, the probability of both the MSP and LSP being lost and the probability of only the LSP being lost for the non-loss sensitive cells are obtained. The waiting time distributions of the delay sensitive and delay insensitive cells are also obtained.

In [60], Arvidsson studies the performance of a circuit switched link fed with calls of different priorities. The allocation of link bandwidth in this case is analogous to the allocation of CLADs considered in [59].

In this paper, the calls are assumed to be accepted at random with probability depending on the priority class of the call and the load on the link at the arrival instant. Assuming exponentially distributed holding times, the steady state probability of losing a call, the moments of the blocking periods, the time dependent probabilities of loss and the moments of the overflow intervals for each class are evaluated. The arrival process of calls has been considered to be either a Poisson process or has been modelled as an MMPP.

In [61], Yong et al use the matrix analytic method to analyze the queueing characteristics of an integrated services TDM system with two priority classes. A TDM frame with a capacity of M packets is used to transmit a high priority packet at every time step. Any unused capacity is utilized to transmit low priority packets. This system is modelled as a discrete time system with a fluctuating number of servers for the low priority packets. The LP sources are assumed to be Poisson. The dependency of the aggregate arrivals from several high priority sources is represented by a one-step transition probability matrix. Under this formulation, the mean queue length and the standard deviation of the queue length for both the classes are obtained.

In [62], Potter et al consider a multi-priority queueing system which involves several distributed local queues and a central server. This priority model is an extension of Kleinrock's processor sharing model [63] with generalization to multiple priorities and the addition of a buffer at each traffic source for each priority class to buffer additional packets queued at that source. Assuming the traffic from each priority class to be modelled as a Poisson process and under a non-preemptive priority discipline, the $M/G/1$ queue is studied to evaluate the mean packet delay as a function of the packet length for each priority class. The application of these results for a DQDB sub network is also considered.

In [64], Khamisy et al consider a class of discrete priority queueing systems with Markov modulated arrivals. In these systems, N queues are served by a single server according to priorities that are assigned to these queues. Packet arrivals are modelled as discrete time batch processes with a distribution that depends on the state of an independent, common, two state Markov chain. Expressions for the moments of the queue lengths as well as for the average delays are obtained. While this formulation is able to cater to applications where the parameters of the arrival processes are not fixed over time, the *commonness* of the underlying modulating process limits these results to be unsuitable for cases where the sources are independent and follow their own modulating processes.

In [65], Chen et al study a packet switch with input queues and two priority classes. The arrivals from both the priority classes are assumed to be modelled as Bernoulli processes. Higher priority packets preempt the service of lower priority packets in the same queue. When a Head of the Line (HOL) contention occurs, higher priority packets are preferred. In case of contention between the high priority packets, two strategies are considered. In one case, the high priority packets which cannot be transmitted in a single attempt are dropped. In another case, these packets are also queued back for further attempts. It is obvious that the high priority traffic sees only a single priority switch. To study the performance for the low priority switch, the switch is viewed as a system of N parallel queues with N servers. If we denote the probability of the low priority packet being successful in a single attempt as p , then the no. of slots required for transmission of a low priority packet has a geometric distribution with parameter p . Under some independence assumptions, the resulting GEO/GEO/1 queue is studied and the queue length distribution of the low priority packets is obtained. The expression

for the maximum throughput of the switch is also derived and is found to exceed that of a single priority switch

In [66], Mitrou et al study the performance of the ATM multiplexer with two service classes and a combined space priority and service priority mechanisms. The server is allotted to the two classes probabilistically with prob p and $1-p$. (It is obvious that for p other than 0.5, one of the classes will get a better service rate). If the cell belonging to a particular class is absent in a time slot, then that slot is given to the other class. The buffer space is partitioned into two and a cell arriving at the buffer when the portion of the buffer belonging to its class is full, is lost. With this formulation, the queueing behaviour can be characterized by a two dimensional Markov chain. Using this, the buffer occupation probabilities as well as the cumulative density function of the queueing delays of cells from each priority class are obtained through numerical analysis. With some approximations, the problem reduces to two loosely coupled GEO/GEO/1 queues and analytic expressions for the buffer occupancy probabilities and queueing delays are obtained.

The work done on the optimum queueing policies for an ATM switch will also be of some relevance to our present work and is briefly reviewed next. Optimum queueing policies for catering to the needs of both delay sensitive and loss sensitive applications are considered in Awater et al [67]. It is shown that these complementary performance requirements can be exploited with an LDOLL (Low delay or low loss) queue where the sources get either service priority or storage priority. Using Markov decision theory and concepts of linear programming, an efficient solution for the LDOLL switch is obtained. It is observed that the performance of a cell type improves, the more this type is in a minority. This effect is most articulate for low loss (LL) cells. In a typical case, the loss probability goes from 10^{-6} with 80% LL cells to 10^{-12} .

with 20% LL cells. On the same range, the delay for the low delay (LD) cells halves and the delay variance shows a corresponding reduction.

In [68], Hiroshi Saito considers the optimum queueing discipline that minimizes the number of cells being discarded when their delays exceed the maximum permissible value. It is shown that the earliest due date discipline (EDD) - alternately known as the dynamic priority discipline - is optimum for minimizing the loss of cells due to outage. Further, the EDD discipline also preserves the order of the cells within a class and resequencing at the receiving node is not necessary. Several versions of the dynamic priority discipline have been discussed in the literature [69-72]. We review two of them in more detail.

Youngho Lim et al [71] proposed and analysed a dynamic priority discipline known as HOL-PJ (Head of line with priority jumps) for a packet switch serving multiple classes of delay sensitive traffic. Several levels of priority are proposed. A packet at the lowest priority level jumps to the next higher priority level if the delay experienced by that packet exceeds a particular threshold. This process is repeated for packets at each priority level. The packet with the largest queueing delay in excess of its delay requirement gets the transmission priority. It is shown that under realistic traffic conditions, for the different classes of traffic, this discipline can make the tail probabilities of the delay distribution in excess of their respective delay requirements approximately the same.

In [72], Fratini considers a dynamic, non-preemptive priority queueing system with a single server and two independent Poisson streams of customers with general service time distributions. Type 2 customers have a finite waiting room of size $K-1$ and Type 1 customers have infinite waiting space. The server offers higher priority to Type 2 customers when there are at least N of

them in the queue, otherwise, the Type 1 customers get higher priority. For this model, the various queueing characteristics of both priority classes are obtained using the matrix geometric approach.

In [73], Hluchyl et al. compares the relative merits of the FCFS, Head of the Line (HOL) and Weighted Round Robin (WRR) disciplines for the integrated fast packet networks. It is shown that with the FCFS discipline, in order to obtain sufficiently small packet loss probability either low link utilization or small source peak rate is required. With the HOL discipline, when the burst sizes and the delay requirements of the various classes are different, then better delay performance for the high priority traffic can be achieved with a marginal increase in the delay for the low priority traffic. However, if the delay requirements of the various classes are the same, the HOL discipline is shown to be unsatisfactory. In this case the WRR discipline is found to be satisfactory as it ensures that none of the priority classes are starved of service. A hybrid WRR and HOL discipline is proposed to combine the advantages of each of these disciplines.

In [74], Hashida et al. present a conservation law for a class of discrete time queues and use it to find the mean waiting times in a SBBP, BBP/D/1 queue with non-preemptive priority discipline. The high, and low priority traffic are modelled as a Switched Batch Bernoulli process and a Batch Bernoulli process, respectively.

In [75], Zhung studies the performance of a dual priority system with two separate queues and a single server using a fluid flow approach. The traffic arriving at both the queues are assumed to be modulated by the state of a common two state Markov chain. Expressions for the mean queue lengths of both the queues as well as the mean waiting times are obtained as a function of the fraction of the traffic offered to the high priority queue, this fraction is

denoted as α . It is found that the average waiting time is more sensitive to α than the average queue lengths.

Gravey et al [38] study the effect of combining loss priority and service priority for two priority classes in an ATM switch with an output buffer. The output buffer of the ATM switch is modelled as a single server queue. Assuming an M/D/1 model with a non-preemptive priority discipline, the LST of the waiting time of both classes of cells as well as the moments of the delays are computed. For implementing space priority, a push out mechanism is assumed. It is found that, by incorporating the service priority, the delay for the high priority cells can be reduced significantly at the expense of a marginal increase in the delay of the low priority cells. When both space and time priorities are used for a particular class of cells, it is found that the loss probability is significantly lower than the case where only loss priority is used. Hence service priority may even be used to improve the loss performance of the switch.

In [76], Gupta et al study the performance of a fast packet switch with a dual plane switch architecture and an input buffer. Assuming an M/D/1 model for the queueing at the input buffer of the switch and assuming two priority classes with a non-preemptive priority discipline, the mean delay for the cells belonging to each priority class can be obtained.

In [77], Jacob et al also study the performance of a packet switch with input buffers. Here, the scenario in which the traffic at the various input ports of the switch have different burstiness in the selection of output port is considered. For example, if the traffic at one port originates from local traffic either from a Metropolitan Area network or from B-ISDN terminals, the resulting traffic will seek all output ports from time to time and the duration for which it seeks a particular output port will be bursty. On the

other hand, when the traffic is a transit traffic from an adjacent ATM node, then the probability that it seeks a particular output port and the duration for which it continues to seek the output port will be high and hence is termed as a "smooth " traffic Output port contention occurs in the switch when a number of input ports look for connections to a single output port It is shown that under this scenario, output port contention in a switch can be minimized and the throughput of the switch can be increased by offering non-preemptive priority for the cell stream with the largest burstiness Assuming two priority classes, the high priority input queue is modelled as a GEO/GI/1 queue and expressions for mean waiting time of a burst is obtained The mean waiting time of a burst from the low priority class is analysed using a more complex Markovian model

In [78], Schormanns et al consider an ATM switch modelled as a GEO/D/1 queue with two priority classes The distribution of the waiting times for the low and high priority cells for both preemptive and non-preemptive service disciplines are obtained using combinatorial methods The server is assumed to be synchronous, i.e the cells that arrive at an empty queue do not receive service immediately but wait for the beginning of the next time slot Extension of these results for the corresponding $\text{GEO}^{[x]}/G/1$ priority system is considered in [79]

Some general limitations of the type of studies described above can also be summarized It can be noted that the queueing analysis with a SBBP/D/1 becomes intractable as the number of phases of the modulating process increases Markov Modulated fluid flow models have the the same problem, moreover, they cannot take care of periodicities in the input Even though performance studies of ATM switches becomes tractable with assumptions of Poisson and/or Bernoulli arrival processes, the results obtained thereby are not always

accurate Neither of these processes are capable of taking into account the periodicities, correlations and burstiness that may occur in the input to an ATM switch Results in [80] have demonstrated that the ATM switch performance is much worse for periodic streams as compared to the Bernoulli cell arrival process While periodicities at the edges of the ATM network arise due to source periodicities, periodicities on links inside the network could arise due to certain aspects of the switch design For example, in the preferred method of output buffering in a switch, the output appears periodically due to the slotted nature of the link Further, in addition to periodicities in the traffic, burstiness may also arise as the link alternates between talk and silence periods (corresponding to non-empty and empty states of the buffer)

2.5. AN OVERVIEW OF THE PROBLEM

In view of the inaccuracies that result from the simple models considered for an ATM network with priorities, it is interesting to know how "good" an MMPP model can be in such a situation Even though an MMPP model cannot take care of periodicities in the input [40], it can take care of correlations and burstiness in the input stream and is computationally simple to use In [50], the performance of the switch obtained using an MMPP model is modified to take care of periodicities and the results are compared with that obtained using a more accurate discrete time model It is found that the MMPP model is reasonably accurate in predicting the performance of the switch

In view of the above and because MMPP has been widely used to study various overflow systems, we choose MMPP to characterize the traffic arriving from each priority class The traffic from each priority class is considered as arriving at separate queues A non-preemptive priority discipline and constant service time/cell from each priority class are also assumed Under

these assumptions, the evaluation of the queue length densities and queueing delays corresponding to each priority class, when the queue capacities are either infinite or finite, is the central problem that is considered in our present work

The methodology used for the analysis given in our present work is summarized briefly next. We generally restrict ourselves to a dual priority and assume that the low and high priority traffic arrive at two separate queues denoted as Q_1 and Q_2 , respectively. The method given in Neuts [81] for the analysis of queues under the framework of structured stochastic matrices of the $M/G/1$ type has wider applicability and has been used to study the queues with more complex arrival processes [81-83]. In fact, the $MMPP/D/1$ queue with nonpreemptive priority has a lot of similarities to an $N/G/1$ queue and hence the approach used by [82] can be adopted appropriately for the present problem.

It is similar to the $N/G/1$ queue because when the busy period starts there may be more than one customer in the respective queues. It may be noted here that when the busy period (BP) of the low priority Q_1 queue starts, there may be more than one customer in Q_1 . The first customer arriving at Q_1 will have to wait for the high priority Q_2 queue to become empty and, in this period more customers may arrive at Q_1 . Even the first customer arriving at an empty Q_2 may have to wait for the ongoing service, if any, for a Q_1 cell to be over before receiving service, hence Q_2 's BP may also start with more than one customer. In that respect, the present problem differs from an $MMPP/D/1$ with FCFS discipline and resembles an $N/G/1$ queue. However it differs from the $N/G/1$ queue in three respects. Firstly, unlike in an $N/G/1$ queue, the time when the queue becomes non-empty and the time when BP starts are not the same. Secondly, the number of customers in the queue (either in Q_1 or Q_2) when the

BP of Q1 or Q2 starts also depends on the state(whether empty or non-empty) of the other queue Thirdly, the inter departure time of the customers from Q1 depends on the phase of the MMPP to Q2 at the previous departure instant Hence the transition probability matrices pertaining to Q1 and Q2, though similar to the one corresponding to N/G/1 queue, are coupled and are more complex

We shall see in the next chapter that we should know the busy period distribution of the server in Q2 in order to study the queues Q1 and Q2 (We develop a recursive procedure for the computation of the busy period distribution of Q1 and Q2 in Chapter 4) The maximum length up to which the distribution of the BP of Q2 needs to be known depends on the traffic offered at Q2 and determines the computational complexity and the storage requirements required for the above study When the higher priority traffic is very high, the storage requirements and computational complexity may become unmanageable when the storage capacity for Q2 is assumed to be infinite In practice, this may not be a serious limitation due to the following reasons Firstly, as observed in [60], if significant reductions in the cell delay and its jitter is desirable, it can be achieved effectively only if the higher priority cells are in a minority Hence we expect that the high priority traffic will be small Secondly, if the high priority traffic is indeed high then the practical buffer sizes used cannot be considered to be infinite and we have to actually deal only with a BP that results from a finite buffer In fact this second reason motivated us to consider the non-preemptive MMPP/D/1/K priority system as a follow up problem

The priority system considered in our present work will be useful in the study of ATM systems incorporating priority service Firstly, it can be used to determine the buffer sizes required for an ATM switch given the maximum

high and low priority traffic it is expected to carry. Secondly, it can be used in CAC to determine whether the new call can be accepted or not. For it to be executed on line, CAC procedures should be simple and less time consuming. In this context, we examined the possibility of reducing the computational and storage complexity by sacrificing some accuracy (after all in CAC one is interested only on the bounds on the QOS parameters and not the exact values). This led to two approximate models which are computationally simple and give correct results under most normal traffic conditions. For the computation of various moments of queueing delays at Q2, we derive an expression for the LST of the virtual queueing delay at Q2. Using this the average queueing delays at Q1 and Q2 are evaluated.

As mentioned earlier the non-preemptive MMPP/D/1/K system is studied along the same lines. The computation of the busy period distribution of the finite sized Q2 is considered by developing an efficient recursive procedure. The validity of the computational approach for all the above cases is verified by comparing the results obtained using the numerical computation with that obtained using simulation. Finally, we consider the extension of our results when the number of priority classes are more than 2.

REFERENCES

- 1 S. E. Minzer, "Broadband ISDN and Asynchronous reansfer Mode", IEEE Communications Magazine, Sep 1989, pp 17-24
- 2 K. Apostolidis, L. F. Merakos and X. h. Xing, "A reservation protocol for packet voice and data integration in unidirectional bus networks" IEEE Trans Commns Mar 1993, pp 478-485
- 3 P. Papantoni-Kazakos, "Multiple access algorithm for a system with mixed traffic High and Low priority", IEEE Transactions on Communication Mar

1992, pp 541-555

- 4 R K Goel and Elkakeem, "A hybrid FARA/CSMA-CD protocol for voice-data integration", Computer Networks and ISDN systems, Mar 1985, pp 223-240
- 5 Chlamtac I, "AN Ethernet compatible protocol for real time voice/data integration", Computer Networks and ISDN systems, 1985, pp 81-96
- 6 Gagan L Choudhury, Stephen S Rappaport, "Priority access schemes using CSMA/CD", IEEE T Commn, July 1985, pp 620-626
- 7 Tobagi, F A, "Carrier sense multiple access with message based priority functions", IEEE Transactions on Communication, Jan 1982, pp 185-200
- 8 Raif O Onvural, "Asynchronous Transfer mode networks Performance issues", Artech House, Boston, 1994
- 9 A G Fraser, "Early experiments with Asynchronous Time Division Networks", IEEE Network, Jan 1993, pp 12-26
- 10 J Y Boudec, "The Asynchronous Transfer Mode, a tutorial", Computer Network and ISDN systems 15, May 1992, pp 279-308
- 11 Saewoong Bahk, M L Zarki, "Routing in ATM Networks", ITC Specialist Seminar Oct 90, Morristown, NJ, USA
- 12 Ahmadi, H, and W E Denzel, "A survey of modern High performance switching techniques", IEEE Journal on SAC, Sep 1989, pp 1091-1103
- 13 Michael G Hluchy and Mark J Karol, "Queueing in high performance Packet switching", IEEE J SAC, Dec 1988, pp 1587-1597
- 14 Ellen Witte Zegura, "Architectures for ATM switching systems", IEEE Commn magazine, Feb 1993, pp 28-37
- 15 Achille Pattavina, "Non blocking architectures for ATM switching systems", IEEE Commns magazine, Feb 1993, pp 38-48
- 16 F A Tobagi, "Fast packet switch architectures for Broadband Integrated Services Digital Networks", Proc IEEE, Vol 78, No 1 Jan 1990, pp 133-167

- 17 C Rasmussen, J Sorensen et al "Source independent Call admission procedures in ATM networks", IEEE Transaction on Selected areas in communication, April 1991, pp 351-358
- 18 E Dutkiewicz and G Anido, "Connection Admission Control in ATM networks", Proc ITC specialist seminar, Kracow 1990, pp 166-176
19. Xiaoqiang Chen, "Modelling Connection Admission Control", Proc. INFOCOM-93, pp 274-281
20. Masayuki Murata, Yuji Oie, T Suda and H Miyahara, "Analysis of a discrete time single server with bursty inputs for traffic control in ATM networks, GLOBECOM'89, 1989, pp 1781-1787
21. Tadanobu Okada, Hirokazu Ohnishi and Naotaka Morita, "Traffic control in Asynchronous Transfer Mode Networks", IEEE Commn Magazine, Sep 1991, pp 58-62
- 22 Ferit Yegenoglu and Bijan Jabbari, "Performance evaluation of MMPP/D/1/K queues for aggregate ATM Traffic models", Proc INFOCOM 93, pp 1314-1319
- 23 D A Hughes, G Anido and H S Bradlow, "Queueing analysis for multiplexed bursty traffic with application to ATM switch performance", Proc ITC specialist seminar, Kracow, pp 93-103
- 24 R Guerin, H Ahmadi and M Naghshineh, "Equivalent capacity and its application to bandwidth allocation in High speed networks", IEEE T SAC, Sep 1991, pp 968-981
- 25 Zbigniew Dziong, Ke Qiang, Liao and Lorne-Masorz, "Buffer dimensioning and Effective BW allocation in ATM based Networks with priorities", ITC specialist seminar, Kracow, pp 154-165
- 26 Kenechi Mase and Shigeo Shido, "Real time Network Management for ATM networks", Proc ITC-13, pp 129-136

- 27 E Wallmier and C M Haubei, "Blocking probability in ATM pipes control led by a call acceptance algorithm based on Mean and Peak rates", Proc ITC-13, pp 137-142
28. J W Roberts, "Variable bit rate traffic control in B-ISDN", IEEE communication Magazine ,Sep 1991, pp 50-56
29. J D. Burgin and D Dorman, "Broadband ISDN resource management The role of virtual paths", IEEE communication Magazine ,Sep 1991 pp 44-48
- 30 E P Rathgab, "Modelling and Performance comparison of policing mechanisms for ATM networks", IEEE Transaction on Selected areas in communication", April 1991 pp 325-334
31. J J. Bae and T Suda, "Survey of traffic control schemes and protocols in ATM networks, Proc IEEE, Feb 1991 pp 170-189
- 32 K Sriram, R. S Mckinney and M H Sherif, "Voice packetization and compression in B-ISDN networks", IEEE Transaction on Selected areas in communication, April 1991, pp 394-404
33. Adrian E Eckberg, Bharat T Doshi and Ricchard Zoccolillo, "Controlling congestion in B-ISDN/ATM Issues and Strategies", IEEE Commn Magazine, Sep 1991, pp 64-70
34. Duke Hong and Tatsuya Suda, "Congestion control and prevention in ATM networks", IEEE Network magazine, July 1991, pp 10-16
- 35 Jean Yves Le Boudec, "An efficient Solution method for Markov models of ATM links with loss priorities, IEEE Trans on Selected Areas in communication, Apr 1991, pp 408-417.
- 36 Hans Kroner, G Huberterne, P Boyer and A Gravey, "Priority management in ATM switching nodes", IEEE Trans on Selected Areas in communication, Apr 1991, pp 418-426.

- 37 A Y M Lin and J Silvester, "Priority queueing strategies and buffer allocation protocols for traffic control at an ATM Integrated Broadband Switching system", IEEE T SAC Sep 1991, pp 1524-1536
- 38 A.Gravey and G Hebutterne, "Mixing time and loss priorities in a single server queue", Proc ITC - 13, Copenhagen, June 1991, pp 147-152.
- 39 Hiroshi Saito, "Queueing analysis of cell loss probability control in ATM networks", Proc ITC-13, Copenhagen, June 1991, pp 19-23
- 40 V. Ramaswamy and W Willinger, "Efficient traffic performance strategies for packet multiplexer", ISDN Networks and Systems, Dec 90, pp 401-407
41. S Kowtha and D R Vaman, "A Generalized ATM Traffic model and its Application in Bandwidth Allocation", Proc of ICC '92, pp 1009-1013
- 42 A Kuczura, "The Interrupted Poisson Process as an overflow process", BSTJ 52, 1987, 437-448
- 43 K. S. Meier and Hellstern, "The analysis of a queue arising in overflow model", IEEE Trans commns , 37(4), 1989, pp 367-372.
- 44 I. Ide, "Superposition of Interrupted Poisson Processes and its application to packetized voice multiplexers", Proc ITC-12, Torino 1988
- 45 H. Heffes and D M Lucantonì," A Markov Modulated characterization of Packet voice and Data traffic and Related Statistical Multiplexer Performance", IEEE Trans on Selected Areas in communication ,No 6, pp 856- 867, Sep 1986
- 46 M F. Neuts, "Matrix-Geometric Solutions in Stochastic Models An Algorithmic approach ", The John Hopkins University Press, 1981
- 47 Baiocchi A et al " Loss performance analysis of an ATM multiplexer loaded with high speed On-OFF sources", IEEE Trans on Selected Areas in communication Vol 9, NO 3, Apr 1991, pp 388-393

- 48 H Saito, M Kawarasaki and Hiroshi Yamada, "An analysis of Statistical multiplexing in an ATM transport network", IEEE Trans on Selected Areas in communication, Apr 1991, pp 359-367
- 49 Nagarajan R , J F Kurose and D Towsley, "Approximation techniques for computing packet loss in Finite buffered voice Multiplexers", IEEE Trans on Selected Areas in communication, Vol 9, NO 3, 1991, pp 368-377.
- 50 V. Ramaswami, M Rumsewicz, W Willinger and T Eliazov, "Comparison of some traffic models for ATM performance studies", Proc ITC-13, 1991, pp 7-12
51. Wolfgang Fischer and Kathleen Meier-Hellstern, "The Markov Modulated Poisson Process(MMPP) cookbook", Performance evaluation (18) 1993 pp 149-171
- 52 Anick D, D Mitra and M M Sondhi, "Stochastic theory of a data handling system with multiple sources," BSTJ Vol 61, No 8, Oct 1982, pp 1871-1894
- 53 B. Bensaou, J Guibert and J W Roberts, "Fluid queueing models for a superposition of on/off sources", Proc ITC specialist seminar, Morristown, Oct 90, paper 9.3
- 54 J. R Louvian, P Boyer and A Gravey, "A discrete time single server queue with Bernoulli arrivals and constant service time", Proc ITC-12, pp 1305-1311
- 55 C. Blondia and O casals, "Statistical multiplexing of VBR sources A matrix analytic approach", Performance evaluation, 1992, pp 5-20
- 56 G. D. Stamoulis, M E Anagnostou, A D Georgantas and E N Protonotariors, "A survey of Traffic source models for ATM networks", Proc ITC specialist seminar, Bangalore, Nov 1993, pp 177-184
- 57 S Q Li, "Generalized solution technique for discrete queueing analysis of multimedia traffic on ATM", IEEE T Commns July 1991, pp 1115-1132

- 58 Sanjay K Bose and Les Berry, "An iterative approach to the analysis of a finite ATM buffer for a source with a given state transition probability matrix", Proc of ITC specialist seminar, Bangalore, 1993, pp 145-152
- 59 Zhisheng Niu and Haruo Akimaru, "Studies on mixed delay and non-delay systems in ATM network", Proc ITC-13, 1991, pp 515-520
- 60 Ake Arvidsson, "On the performance of a circuit switched link with priorities", IEEE Journal on Selected Areas in communication No 9 pp 212-219, Feb 1991
- 61 Oliver W Yong and Jon W Mark, "Queueing analysis of an integrated services TDM system using a Matrix analytic method", IEEE Journal on Selected Areas in communication, Vol 9, No 1, pp 88-94, Jan 1991
- 62 Phillip G Potter and Moshe Zukerman, "Analysis of a discrete Multipriority queueing system involving a central shared processor serving many local queues", IEEE Journal on Selected Areas in communication No 9 pp 194-202, Feb 1991
- 63 L Kleinrock, "Queueing systems Vol 2 Computer applications", Wiley, New York, 1976
- 64 Asad Khamisy and Moshe Sidi, " Discrete time priority queues with 2 state Markov Modulated arrivals " Commun-Statist-Stochastic Models 8(2), 1992, pp 337-357
- 65 J S C Chen and R Guerin, "Performance study of an input queueing packet switch with two priority classes", IEEE Trans Commns Jan 1991, pp 117-126
- 66 W M Mitrou and D E Pendarakis, "Cell level statistical multiplexing in ATM networks Analysis, Dimensioning and Call acceptance control w r t QOS criteria", Proc ITC-13, pp 7-12

- 67 Geert A Awater and Frits C Schoute, "Optimal Queueing policies for fast packet switching of Mixed traffic" IEEE Journal on Selected Areas in communication No 9, pp 458-467, April 1991
- 68 Hiroshi Saito, "Optimal Queueing discipline for real time traffic at ATM switching nodes" IEEE Transactions on Communication pp 2131-2136 Dec 1990
- 69 R W Conway, W L Maxwell, L W Miller, "Theory of scheduling ", Addison Wesley, USA, 1967
- 70 N K Jaiswal, "Priority Queues", Academic press, New York, 1968
- 71 Youngho Lim and John E Kobza, "Analysis of a delay dependent priority discipline in an Integrated multiclass traffic fast packet switch" IEEE T commns pp 659-665, May 1990
- 72 Stephen S Fratin, "Analysis of a dynamic priority queue", Commun-Statist-Stochastic Models 6(3), 1990, pp 415-444
- 73 M G Hluchyl and A Bhargava, "Queueing disciplines for integrated fast packet networks", Proc of ICC '92, pp 990-996
- 74 On Hashida and Yooshitaka Takahashi, "A discrete time priority queue with switched Batch Bernoulli process inputs and constant service time", Proc ITC-13, pp 521-526
- 75 Ji Zhung, "Performance study of Markov Modulated Fluid Flow models with priority traffic" Proc INFOCOM-93, pp 10-17
- 76 Anil K Gupta and N D Georganos, "Priority performance of ATM packet switches", Proc Infocom-92, 5D 2 1 (1992)
- 77 Lilly K Jacob and Anurag Kumar, "Comparative performance of scheduling strategies for switching and Multiplexing in a Hub-based ATM network A simulation study" ITC specialist seminar Bangalore, Nov 1993, pp 35-44
- 78 J A Schormanns, Jonathan M Pitts, E M Scharf, "Priorities in ATM switches", ITC-13 (1991)

- 79 J A Schormanns, Jonathan M Pitts, E M Scharf, " A priority queue with superimposed Geometric batch arrivals", Commun-Statist-Stochastic Models 9(1), 1993, 105-122
- 80 T E Eliazov, V Ramaswami, W Willinger and G Latouche, "Performance of an ATM switch Simulation study", IEEE Proc of INFOCOM 90, 1990, pp 644-659
81. M F Neuts, "Structured stochastic matrices of the M/G/1 type and,their application", Marcel Dekker, 1989
- 82 V Ramaswamy," The N/G/1 queue and its detailed analysis," Adv Appl Prob , Vol 12 pp 222-261, Mar 1980
- 83 Lucantonì D M, "New results on a single server queue with a batch Markovian arrival process", Commun-Statist-Stochastic Models 7(1), 1991, pp 1-46

CHAPTER 3

EMBEDDED SEMI-MARKOV SEQUENCES AND TRANSITION PROBABILITY MATRICES OF THE HIGH AND LOW PRIORITY QUEUES

3.1. INTRODUCTION

The queueing analysis of an MMPP/D/1 non-preemptive priority system with two priority classes and constant service time/customer from each priority class is considered in this thesis. A possible area of application for these results will be in an ATM network for the design of the ATM nodal buffer as well as for use in Call Admission Control (CAC) when non-preemptive priority service is being provided. As mentioned in Chapter 2, we adopt a numerical approach for obtaining the characteristics of the prioritized system. The evolution of the states of the high and low priority queues are stochastic processes. For studying the statistical characteristics of these states, we have to first choose appropriate embedded points in the time evolution of these processes which satisfy the Markovian properties. Having thus chosen the semi-Markov chains (SMC), the transition probability matrices of these SMCs are quantified. Once these matrices are known, the evaluation of the statistics of the queues can be done using numerical analysis, as in Kleinrock[1]. We consider the selection of the embedded points and the evaluation of the transition probability matrices.

The basic unit of information transferred from one node to another, corresponding to a particular connection (call), is referred to as "cell" in ATM terminology. This corresponds to the "customer" in a queueing situation. We therefore use the terms "customer" and cell interchangeably in present work. For ease of analysis, we assume that the delay sensitive and non-delay sensitive cells arrive at two separate queues Q1 and Q2 as shown in Fig. 3.1. A single server is shared between the two queues on a non-preemptive

basis. The cells arriving at Q2 have higher priority than those arriving at Q1 and the priority is assumed to be incorporated at the call level. All the cells belonging to a particular call arrive at the same queue (either Q1 or Q2) and are served with the same priority. The server is assumed to be asynchronous i.e. a cell arriving at either Q1 or Q2 when the server is idle, starts receiving service immediately. All cells require a constant service time of D sec/cell, irrespective of their priorities.

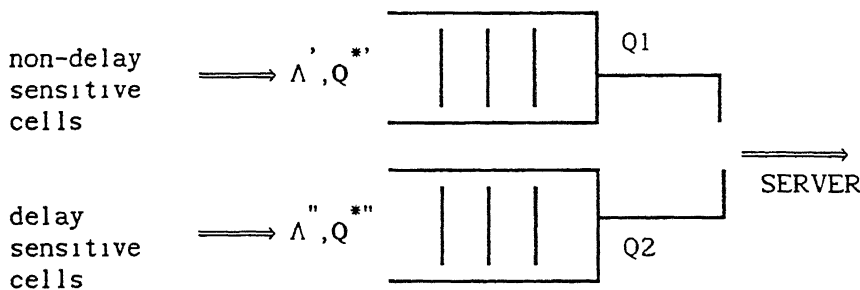


Fig 31 Model for the time priority mechanism

The cells arriving at Q1 is assumed to occur according to an M-phase Markov modulated Poisson Process(MMPP) with arrival rate and infinitesimal generator matrices given by Λ' and $Q^{*'}$ respectively. Similarly, the arrivals to Q2 occur according to an N phase MMPP with arrival rate and infinitesimal matrices given by Λ'' and $Q^{*''}$ respectively.

Under these assumptions, the probability distribution of the occupancy of the queues Q1 and Q2 as well as the delays experienced by the cells arriving at these queues are evaluated using the Matrix-Analytic approach of Neuts [2], [3]. In the next section, we give more details about this approach. In essence, by choosing the embedded points appropriately, we obtain two Semi-Markov Sequences (SMCs) - one corresponding to each queue. Invariant probability vectors of the transition probability matrices corresponding to

these SMCs give the required queue occupancy probabilities

It can be noted that these vectors cannot be independently obtained as the transition probability matrices of Q1 and Q2 are coupled. This can be verified as follows. In view of the non-preemptive priority used by the server, the first cell arriving at an empty Q2 does not always receive service immediately. If Q1 is non-empty, it waits for the Q1 cell under service to complete its service, during this waiting time more cells may arrive at Q2. Hence the number of cells in Q2 when the busy period (BP) of the server in Q2 starts and the duration of the (BP) depends on the probability of Q1 being empty at an arbitrary time instant. Further, at the departure instants of the cells from Q1, if Q2 is non-empty, the server goes on "vacation" to clear the cells in Q2. The probability of occurrence of this vacation period (BP of the server in Q2) as well as its duration depends on the characteristics of the arrival process to Q2. Hence the inter-departure times of cells from Q1 and hence the transition probability matrix of Q1 depends on the BP distribution of Q2 as well as on the probability of Q2 being empty. We noted already that the BP of Q2 depends on the probability of Q1 being empty. Hence the transition probability matrices are coupled. However, the desired queue length densities (ie the probability that the queue length is n , $n \geq 0$) can be evaluated iteratively.

The following notational conventions are used throughout the thesis. All the parameters pertaining to Q1 are indicated by a superscript of (') and those of Q2 by ("). Matrix mass functions of time are denoted by bold faced, upper case alphabets and the Laplace Steiltjes Transform (LST) of these functions are denoted by a superscript of tilde (~). The vectors are denoted by bold faced, lower case alphabets.

3.2. EMBEDDED SEMI - MARKOV SEQUENCES OF Q1 AND Q2

We denote the departure epochs of cells from Q1 and Q2 as τ'_n and τ''_n respectively for $n = 1, 2, 3, \dots$. The number of cells in Q1 (ie both those waiting and the one possibly being served) and the phases of MMPP1 and MMPP2 at τ'_n are denoted by X'_n , $J_n^{(1)'}$ and $J_n^{(2)'}$ respectively. At an arbitrary time t , these parameters are denoted by $X'(t)$, $J^{(1)'}(t)$ and $J^{(2)'}(t)$, respectively. Similarly, the number of cells in Q2 (those waiting and being served), the phase of the MMPP1 and MMPP2 at τ''_n are denoted as X''_n , $J_n^{(1)''}$ and $J_n^{(2)''}$ respectively, these parameters at an arbitrary time t are denoted as $X''(t)$, $J^{(1)''}(t)$ and $J^{(2)''}(t)$ respectively. It can be noted that, conditioned on the phases of the MMPPs at τ'_{n-1} , the successive inter departure times from Q1, $(\tau'_n - \tau'_{n-1})$ for $n=1, 2, 3$ are independent and are identically distributed when $X'_{n-1} > 0$. Hence the sequence $\{X'_n, J_n^{(1)'}, J_n^{(2)'}, \tau'_n - \tau'_{n-1} \mid n \geq 1\}$ forms a Semi-Markov Chain (SMC) with the state space $\{0, 1\} \times \{1, 2, M\} \times \{1, 2, N\}$. Similarly, it can be shown that $\{X''_n, J_n^{(1)''}, J_n^{(2)''}, \tau''_n - \tau''_{n-1} \mid n \geq 1\}$ also forms an SMC with the same state space.

With this formulation, in order to specify the various elements of the transition probability matrices corresponding to these SMCs, six indices are required. However, a compact notation can be obtained by redefining the state variables of the queueing system as follows. Consider the process obtained by superposing the Markov processes (Q^*, Q'') governing the transitions of the states of the MMPPs to Q1 and Q2. Proceeding as in Neuts [3, pp 277] it can be shown that the composite process is a Markov process with the infinitesimal generator \underline{Q}^* given by

$$\underline{Q}^* = Q^* \otimes I_N + I_M \otimes Q'' \quad (3.2.1)$$

where I_N is the $N \times N$ identity matrix and \otimes is the Kronecker product. The Kronecker product of two matrices (see for e.g. Bellman [4]) B and C is

defined as

$$B \otimes C = \begin{bmatrix} B_{11}C & B_{12}C & & B_{1m}C \\ & & & \\ & & & \\ B_{n1}C & B_{n2}C & & B_{nm}C \end{bmatrix}$$

Here B is a $n \times m$ matrix with the $(i,j)^{th}$ elements as B_{ij} . It may be noted that the dimensions of Q^{**} and Q^{**} are M and N respectively. If $J(t)$ is the phase (state) of the composite process at time t , then there exists a unique one-to-one mapping from the phases of the MMPP1 and MMPP2 to that of the composite process and vice versa and is given by

$$J(t) = (M-1)J'(t) + J''(t) \quad (3.2.2)$$

$$J'(t) = J(t) \bmod M \quad (3.2.3)$$

$$J''(t) = \{J(t) - J'(t)\} / (M-1) \quad (3.2.4)$$

In terms of the composite process, we can consider the arrival processes to Q_1 and Q_2 to be from two MN -phase MMPPs denoted as MMPP 1 and MMPP 2, respectively, whose transition rate processes are perfectly correlated and are identical to that of the composite phase process, the phase of the composite process and those of MMPP 1 and MMPP 2 are equal at all time instants. The arrival rate matrices of MMPP 1 and MMPP 2 are chosen as follows. Since the arrivals to Q_1 depends only on the phase of MMPP 1, the arrival rate matrix of MMPP 1 denoted as $\underline{\Lambda}'$ is also made to depend on only the phase of MMPP 1 as follows

$$\underline{\Lambda}' = \Lambda' \otimes I_N \quad (3.2.5)$$

Similarly, the arrival rate matrix of MMPP 2 denoted as $\underline{\Lambda}''$ is chosen as

$$\underline{\Lambda}'' = I_M \otimes \Lambda'' \quad (3.2.6)$$

where I_M , I_N are $M \times M$ and $N \times N$ identity matrices, respectively. Let λ'_1 and λ''_1 denote the arrival rate of MMPP1, MMPP2 in phase 1, then for $M = 2$ and $N = 2$

$\underline{\Lambda}'$ and $\underline{\Lambda}''$ are given by -

$$\underline{\Lambda}' = \Lambda' \otimes I_N = \begin{bmatrix} \lambda'_1 & 0 & 0 & 0 \\ 0 & \lambda'_1 & 0 & 0 \\ 0 & 0 & \lambda'_2 & 0 \\ 0 & 0 & 0 & \lambda'_2 \end{bmatrix}$$

$$\underline{\Lambda}'' = I_M \otimes \Lambda'' = \begin{bmatrix} \lambda''_1 & 0 & 0 & 0 \\ 0 & \lambda''_2 & 0 & 0 \\ 0 & 0 & \lambda''_1 & 0 \\ 0 & 0 & 0 & \lambda''_2 \end{bmatrix}$$

Defining the composite process, MMPP $\underline{1}$ and MMPP $\underline{2}$ in this fashion, we need to use only one index to keep track of the phases of MMPP 1 and MMPP 2 simultaneously. Let, $\underline{J}'_n, \underline{J}''_n$ denote the phase of the composite process at τ'_n, τ''_n , respectively. Then, it can be observed that the sequences -

$$\{X''_n, \underline{J}''_n, \tau''_n - \tau''_{n-1}, n \geq 1\} \text{ and } \{X'_n, \underline{J}'_n, \tau'_n - \tau'_{n-1}, n \geq 1\}$$

also form SMCs with the state space $[0,1, \dots] \times [1,2, \dots, MN]$. The transition probability matrices of these SMCs pertaining to Q1 and Q2 are denoted as $Q'(t)$ and $Q''(t)$ respectively.

3.3. STRUCTURE OF THE TRANSITION PROBABILITY MATRICES OF Q1 AND Q2

In this section we show that the transition probability matrices $Q'(t)$ and $Q''(t)$ can be expressed in a form similar to that of the transition probability matrices of the queues of the "M/G/1 type" considered in Neuts [3]. Expressing the transition probability matrices of the prioritized queues in the framework of the transition probability matrices of the queues of the "M/G/1 type" require several generalizations. These generalizations are also

useful for the study of queues whose service time distribution depends on the phase of the arrival process

In order to highlight the generalizations that are proposed here, we first consider the structure of the transition probability matrix of the N/G/1 queue. The N/G/1 queue with FCFS discipline is an example of the queues of the "M/G/1 type" and is studied in detail in Ramaswami [5]. Let the departure epochs of customers from this queue be denoted as τ_n for $n = 1, 2, 3$. Let X_n and J_n denote respectively the number of customers in the system and the phase of the arrival process at τ_n . The phase of the arrival process at an arbitrary time instant t , is denoted as $J(t)$. The sequence $\{(X_n, J_n), \tau_n - \tau_{n-1} \mid n \geq 1\}$ forms a Semi-Markov Chain (SMC) with the state space $\{0, 1, \dots, M\} \times \{1, 2, \dots, M\}$ where M is the total number of phases of the arrival process. The transition probability matrix $Q(t)$ of this SMC is given by

$$Q(t) = \begin{bmatrix} B_0(t) & B_1(t) & B_2(t) \\ A_0(t) & A_1(t) & A_2(t) \\ 0 & A_0(t) & A_1(t) \\ 0 & 0 & A_0(t) \\ 0 & 0 & 0 \end{bmatrix} \quad (3.3.1)$$

where $B_m(t)$ and $A_m(t)$ are $M \times M$ matrices with $(i, j)^{th}$ elements given by-

$[A_m(t)]_{ij} = P[\text{Given a departure at time } 0 \text{ which left at least one customer in the system and the arrival process in phase } i, \text{ the next departure occurs no later than time } t \text{ with the arrival process in phase } j, \text{ and during that service there were } m \text{ arrivals}]$

$[B_m(t)]_{ij} = P[\text{Given a departure at time } 0, \text{ which left the system empty and the arrival process in phase } i, \text{ the next departure occurs no later than time } t \text{ with the arrival process in phase } j, \text{ and during that service there were } m \text{ arrivals}]$

than time t with the arrival process in phase j , leaving m customers in the system)

Queues with embedded Markov renewal processes whose transition probability matrices have the above structure are referred to in Neuts [3] as queues of the "M/G/1 type" or queues of the "M/G/1 paradigm"

In order to consider the queue in more detail, we consider the expressions for $B_m(t)$ and $A_m(t)$. Let the service time distribution of the customers be denoted as $H(t)$. Let the counting function associated with the arrival process be denoted as $P(m,t)$, with the $(i,j)^{th}$ elements given by -

$$[P(n,t)]_{ij} = P[N(t)=n, J(t)=j \mid N(0)=0, J(0)=i]$$

$N(t)$ No. of arrivals at the queue in $(0,t]$

Let $U_k(t)$ denote $M \times M$ matrix mass functions with the $(i,j)^{th}$ elements as-

$$[U_k(t)]_{ij} = P[\text{the first batch of customers of size } k \text{ arrive at or before time } t, J(t) = j \mid \text{the queue is empty at time } 0 \text{ and } J(0) = i]$$

Using these matrices, $A_m(t)$ and $B_m(t)$ are given by-

$$A_m(t) = \int_0^t dH(\sigma) P(m,\sigma) \quad m \geq 0, t \geq 0 \quad (3.3.2)$$

$$B_m(t) = \sum_{k=1}^{m+1} \int_0^t dU_k(t-\sigma) A_{m-k+1}(\sigma) \quad (3.3.3)$$

Further generalizations are required if this approach is followed for studying prioritized queues with non-preemptive service discipline of the type required in our thesis. We assume that customers belonging to different priority classes arrive at different queues. Even though, only a dual priority system is considered in detail in this thesis, the generalizations are the same for any number of priority classes and hence we consider that situation here. (Analysis of a Queueing system with more than two priority classes are considered in Chapter 8)

First, let us consider the case when the i^{th} priority queue is non-empty at the time of the previous departure from this queue. Let us compare the distribution of the inter departure time of customers (IDT) from this queue with that of a queue with single priority level (i.e. with FCFS discipline). With the FCFS discipline, the distribution of the IDT from the queue and the distribution of the service time are equal. In a multi-priority system, these distributions are not the same for the low priority queues. This can be verified as follows. The IDT from a low priority queue consists of two parts (i) service time for a customer from this queue and (ii) the busy periods of the higher priority queues that might be initiated by the higher priority customers who arrive during this service. Obviously, the service time distributions take care of only the contribution from part (i). Hence in a multi-priority system the distribution of IDT from the i^{th} priority queue should be used to characterize the transition probability matrix of the corresponding queue.

In the queues of the M/G/1 type, distribution of the IDT from the queue is assumed to be a scalar function and is independent of the phase of the arrival process. In a multi-priority system, IDT from a low priority queue does depend on the phase of the arrival process. In this system, from the point of view of the low priority queue, the server goes on vacation after serving a low priority customer if any higher priority customer is found waiting when the low priority service terminates. This vacation period ends only when all the high priority queues become empty and its duration depends on the arrival rates at these queues. The arrival rates will in turn depend on the phases of the arrival processes to these queues. In view of the dependence of the IDT on the phase of the arrival phase, IDT should be defined as a vector. However, in order to simplify the notation and to retain the form of

the equations for A_m similar to that for an M/G/1 type of queue, we define them as diagonal matrices. It should be noted that the arrivals at the higher priority queue are assumed to be more complex than a simple Poisson process. If this is not true, then IDT can be considered to be a scalar function.

We now consider the IDT corresponding to the case in which the system became empty at the previous departure instant. In the M/G/1 type of queues, the moment a batch of customers (of batch size ≥ 1) arrive at the empty queue, the busy period starts. The time that elapses between this arrival instant and the time when the queue becomes empty once again constitutes the busy period of the server. In a multi-priority system, the arrival instant of the first batch of customers at the i^{th} priority queue need not be the beginning of the busy period of the server. The server might already be busy serving a customer of some other queue. In addition to this, even when the i^{th} priority queue actually becomes empty, the server may still have to continue serving customers of other priority classes. In this thesis, we define the busy period of the i^{th} priority queue as the time that elapses between the beginning of service for the first customer arriving at the i^{th} queue and the time when the queue becomes empty again. Obviously this is different from the busy period of the server.

In a single priority system, the characteristics of the busy period of the server is used to obtain the characteristics of the queue. In the multi-priority system considered in this thesis, we find the busy periods of the queues (defined as above) to be playing a similar role. It may be noted that even when the higher priority arrivals are modelled as Poisson process, the busy periods have to be defined for the individual queues rather than for the server.

For the multi-priority system, we shall choose the argument for the

function " $U_k()$ " to be the time when the busy period of the i th queue starts. It may be recalled that in the queues of the "M/G/1 type", this is chosen to be the time when the first batch of customers arrive at the empty queue.

With these generalizations, We shall show that $Q'(t)$, the transition probability matrix of Q_1 is given by-

$$Q'(t) = \begin{bmatrix} B'_0(t) & B'_1(t) & B'_2(t) \\ A'_0(t) & A'_1(t) & A'_2(t) \\ 0 & A'_0(t) & A'_1(t) \\ 0 & 0 & A'_0(t) \\ 0 & 0 & 0 \end{bmatrix} \quad (3.3.4)$$

where $B'_m(t)$ and $A'_m(t)$ are $MN \times MN$ matrices with (i,j) th elements given by-

$[A'_m(t)]_{ij} = P\{\text{Given that a cell departed from } Q_1 \text{ at time } 0, \text{ leaving at least one cell in } Q_1 \text{ and the arrival process MMPP } \underline{1} \text{ in phase } i, \text{ the next departure occurs at no later than time } t \text{ with MMPP } \underline{1} \text{ in phase } j, \text{ and in the intervening period there were } m \text{ arrivals}\}$

$[B'_m(t)]_{ij} = P\{\text{Given that a cell departed from } Q_1 \text{ at time } 0 \text{ leaving } Q_1 \text{ empty and the arrival process MMPP } \underline{1} \text{ in phase } i, \text{ the next departure occurs at time no later than } t \text{ with MMPP } \underline{1} \text{ in phase } j, \text{ leaving } m \text{ cells in } Q_1\}$

These matrices in turn can be expressed in terms of the $MN \times MN$ matrices $P'(m,t)$, $H'(t)$ and $U'_k(t)$ whose (i,j) th elements are defined by -

$$[P'(n,t)]_{ij} = P[N'(t)=n, \underline{j}(t)=j \mid N'(0)=0, \underline{j}(0)=i]$$

$N'(t)$ No. of arrivals at Q_1 in $(0,t]$

$$\begin{aligned}
[H'(t)]_{1,j} & P[(\tau'_n - \tau'_{n-1}) \leq t \mid \underline{j}'_{n-1} = 1 \text{ and } X'_{n-1} > 0] \delta_{1,j} \\
[U'_k(t)]_{1,j} & P[\text{Busy period of Q1 starts at or before time } t, k \text{ cells} \\
& \text{arrive at Q1 in } (0, t], \underline{j}(t) = j \mid X'(0)=0, \underline{j}(0)=1]
\end{aligned}$$

Using these matrices we next show that $Q'(t)$ is given by (3.3.4) and the matrices $A'_m(t)$ and $B'_m(t)$ are given by

$$A'_m(t) = \int_0^t dH'(\sigma) P'(m, \sigma) \quad m \geq 0, \quad t \geq 0 \quad (3.3.5)$$

$$B'_m(t) = \sum_{k=1}^{m+1} U'_k(t-D) P'(m-k+1, D) u(t-D) \quad (3.3.6)$$

where $u(t)$ is the unit step function

In order to do this, we start with the defining equations for the elements of $Q'(t)$ and consider separately the cases corresponding to $X'_{n-1} > 0$ and $X'_{n-1} = 0$, where the $(1, j, \iota, \not{j})^{\text{th}}$ element of $Q'(t)$ for $\iota \geq 0, \not{j} \geq 0, 1 \leq j, \not{j} \leq MN$ is given by -

$$[Q'(t)]_{1, j, \iota, \not{j}} = P\{X'_n = \iota, \underline{j}'_n = \not{j}, \tau'_n - \tau'_{n-1} \leq t \mid X'_{n-1} = 1, \underline{j}'_{n-1} = j\} \quad (3.3.7)$$

Case 1 $X'_{n-1} > 0$ (Q1 not empty at $(n-1)^{\text{th}}$ departure instant from Q1)

The event $\{X'_n = \iota, \underline{j}'_n = \not{j} \mid X'_{n-1} = 1, \underline{j}'_{n-1} = j, \tau'_n - \tau'_{n-1} = t\}$ implies $(i-i+1)$ arrivals at Q1 from MMPP \underline{j} and a phase transition of j to \not{j} in an interval of t sec. It may be noted that the MMPP is a special case of the versatile point process considered in Neuts [6]. Hence as in Neuts [3, pp 281], we associate with MMPP \underline{j} the $MN \times MN$ probability matrices $P'(m, t)$, whose $(1, j)^{\text{th}}$ elements are given by -

$$\begin{aligned}
[P'(m, t)]_{1,j} &= P\{N'(t) = m, \underline{j}'(t) = j \mid N'(0) = 0, \underline{j}'(0) = 1\} \\
&= 0 \text{ for } m < 0
\end{aligned} \quad (3.3.8)$$

Conditioning on the length of the interdeparture time $(\tau'_n - \tau'_{n-1})$ and using (3.3.8) in (3.3.7) we get -

$$[Q'(t)]_{1,j,\iota,j} = \int_{\sigma=0}^t dH'_{jj}(\sigma) P'_{jj}(\iota-1+1,\sigma) \quad (3.3.9)$$

Writing this in matrix form, we get -

$$[Q'(t)]_{1,\iota} = \int_{\sigma=0}^t dH'(\sigma) P'(\iota-1+1,\sigma) \quad (3.3.10)$$

where $P'(m,t) = 0$ for $m < 0$, $[Q'(t)]_{1,\iota} = 0$ for $\iota < 1-1$. Keeping $\iota=1$ and substituting $\iota = 0,1,2$ in (3.3.10), we get the elements in the second row of $Q'(t)$ as follows

$$[Q'(t)]_{10} = \int_{\sigma=0}^t dH'(\sigma) P'(0,\sigma) = A'_0(t) \quad (3.3.11)$$

$$[Q'(t)]_{11} = \int_{\sigma=0}^t dH'(\sigma) P'(1,\sigma) = A'_1(t) \quad (3.3.12)$$

$$[Q'(t)]_{1m} = \int_{\sigma=0}^t dH'(\sigma) P'(m,\sigma) = A'_m(t) \quad (3.3.13)$$

It can be verified that for $\iota=3,4,5$, the ι^{th} row can be obtained by cyclic shift of the elements of the $(\iota-1)^{\text{th}}$ row of matrices with the leading element replaced by 0, the $MN \times MN$ null matrix

Case 2 $X'_{n-1} = 0$ (Q1 empty at $(n-1)^{\text{th}}$ departure instant from Q1)

In this case the server goes on vacation at τ'_{n-1} and the busy period of Q1 starts again only when Q2 is empty and at least one cell arrives at Q1. Hence $(\tau'_n - \tau'_{n-1})$ is the sum of the vacation period of the server and the service time of one cell. Let the number of cells arriving at Q1 during the vacation interval and the subsequent cell service time be k and $m-k+1$, respectively as shown in Fig 3.2

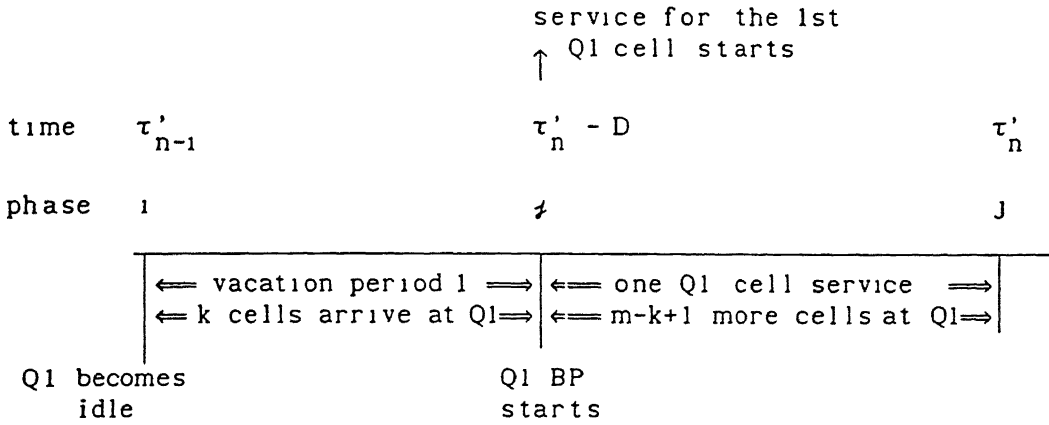


Fig 3.2 Arrivals and phase transitions at Q1 during the IDT of Q1
given that Q1 empty at the last departure instant

Let the phase of the MMPP $\underline{1}$ at τ'_{n-1} , $\tau'_n - D$ and τ'_n be 1, \neq and J, respectively. Then using the definition of $U'_k(t)$ and $P'(m, t)$ along with Fig 3.2, it can be shown that

$$[Q'(t)]_{0,1,m,J} = \sum_{k=1}^{m+1} \int_{\sigma=0}^{t-D} dU'_{k,1\neq}(\sigma) P'_{\neq J}(m-k+1, D) u(t-D) \quad (3.3.14)$$

Integrating (3.3.14) and writing in matrix form we get

$$[Q'(t)]_{0,m} = \sum_{k=1}^{m+1} U'_k(t-D) P'(m-k+1, D) u(t-D) = B'_m(t) \quad (3.3.15)$$

Substituting $m=0,1,2$ it can be verified that we get the elements in the first row of $Q'(t)$. The above derivation shows that $Q'(t)$ indeed has the form given in (3.3.4) with its components obtainable using (3.3.5) and (3.3.6) and the other associated expressions.

The other matrix $Q''(t)$ for the higher priority queue can be obtained in a similar fashion. The $(i, j, \iota, \neq)^{th}$ element of $Q''(t)$ for $i \geq 0$, $\iota \geq 0$, $1 \leq j$, $\neq \leq MN$ is given by -

$$[Q''(t)]_{1,j,\iota,\neq} = P\{X''_n = \iota, \underline{J}''_n = \neq, \tau''_n - \tau''_{n-1} \leq t | X''_{n-1} = 1, \underline{J}''_{n-1} = j\} \quad (3.3.16)$$

Comparing this with (3.3.7), it is obvious that the structure of $Q''(t)$ is the

same as that of $Q'(t)$. Hence, $Q''(t)$ can be obtained by replacing the parameters of $Q1$ by those of $Q2$. Analogous to $Q1$, we define $MN \times MN$ matrices $P''(m,t)$, $H''(t)$ and $U_k''(t)$ whose (i,j) th elements are defined as follows

$$\begin{aligned} [P''(n,t)]_{1,j} & P[N''(t)=n, \underline{J}(t)=j \mid N''(0)=0, \underline{J}(0)=1] \\ N''(t) & \text{No. of arrivals at } Q2 \text{ in } (0,t] \\ [H''(t)]_{1,j} & P[(\tau_n'' - \tau_{n-1}'') \leq t, \mid \underline{J}_{n-1}'' = 1 \text{ and } X_{n-1}'' > 0] \delta_{1,j} \\ [U_k''(t)]_{1,j} & P[\text{Busy period of } Q2 \text{ starts at or before time } t, k \text{ cells} \\ & \text{arrive at } Q2 \text{ in } (0,t], \underline{J}(t) = j \mid X''(0)=0, \underline{J}(0)=1] \end{aligned}$$

$A_m''(t)$ and $B_m''(t)$ can then be expressed as follows

$$A_m''(t) = \int_0^t dH''(\sigma) P''(m,\sigma) \quad m \geq 0, t \geq 0 \quad (3.3.17)$$

$$B_m''(t) = \sum_{k=1}^{m+1} U_k''(t-D) P''(m-k+1,D) u(t-D) \quad (3.3.18)$$

where

$$[A_m''(t)]_{1,j} = P\{\text{Given that a cell departed from } Q2 \text{ at time } 0, \text{ leaving at least one cell in } Q2 \text{ and the arrival process MMPP } \underline{2} \text{ in phase } 1, \text{ the next departure occurs at no later than time } t \text{ with MMPP } \underline{2} \text{ in phase } j, \text{ and in the intervening period there were } m \text{ arrivals}\}$$

$$[B_m''(t)]_{1,j} = P\{\text{Given that a cell departed from } Q2 \text{ at time } 0, \text{ leaving } Q2 \text{ empty and the arrival process MMPP } \underline{2} \text{ in phase } 1, \text{ the next departure occurs at no later than time } t \text{ with MMPP } \underline{2} \text{ in phase } j, \text{ leaving } m \text{ cells in } Q2\}$$

3.4. COMPUTATION OF $A_m''(t)$ AND $U_k''(t)$

For the computation of $A_m''(t)$ using (3.3.17), $P''(m,t)$ and $dH''(t)$ are to be evaluated first. In Sec. 4.4, we give a recursive procedure for the computation of $P''(m,D)$. We next consider the computation of $dH''(t)$. In view of

the constant service time demand of D sec/cell and the high priority of Q_2 , the inter departure time of cells from Q_2 viz $(\tau_n'' - \tau_{n-1}'')$ for $X_{n-1}'' > 0$ is constant and is independent of the phase of MMPP $\underline{2}$ at τ_{n-1}'' . Hence we get -

$$H''(t) = u(t-D)I_{MN} \quad (3.4.1)$$

Using (3.4.1) in (3.3.17), we get

$$A_m''(t) = P''(m,D)u(t-D) \quad (3.4.2)$$

Next, an expression for $\frac{d}{dt}[U_k''(t)]$ is obtained. As mentioned earlier, $U_{1j}''(t)$, the $(1,j)^{th}$ entry of $U_k''(t)$, gives the probability that the busy period of Q_2 starts with MMPP $\underline{2}$ in phase j and k cells arrive at Q_2 at or before time t , given that the idle period started at time 0 in phase 1. Let t be a particular instant when the busy period starts. If Q_1 is not empty (n.e.) when the first cell arrives at an empty Q_2 , it waits for the residual service time (RST) for the current Q_1 cell in service to be over before receiving service and $k-1$ additional cells arrive at Q_2 during this RST. The residual service time of a Q_1 cell is uniformly distributed in the interval $(0,D)$ for $t \geq D$. When t , the time at which the busy period of Q_2 starts is less than D sec, the maximum RST seen by the first cell arriving at an empty Q_2 is t sec. Hence the p.d.f. of RST is given by

$$P(RST=u) = \frac{1}{D} \quad \text{for } 0 \leq u \leq Du(t-D_-) + tu(D-t) \\ = 0 \quad \text{otherwise} \quad (3.4.3)$$

Let the phase of the arrival process at times 0, $t-u$ and t be 1, ℓ and j , respectively. Then

$$\left. \frac{d}{dt} U_{1j}''(t) \right|_{\substack{RST=u, \\ t=t, \\ Q_1 \text{ n.e.}}} = \sum_{\ell=1}^{MN} P_{1\ell}''(0, t-u) \Lambda_{\ell\ell}'' du P_{\ell j}''(k-1, u) \quad (3.4.4)$$

This can be written in matrix form as

$$\left. \frac{d}{dt} U_k''(t) \right|_{\substack{\text{RST}=\omega, \\ t=t, Q1 \text{ n e}}} = P''(0, t-\omega) \underline{\Lambda}'' d\omega P''(k-1, \omega) \quad (3.4.5)$$

Removing the condition on RST using (3.4.3)

$$\left. \frac{d}{dt} U_k''(t) \right|_{\substack{t=t, \\ Q1 \text{ n e}}} = \frac{1}{D} \int_0^{Du(t-D_-)+tu(D-t)} P''(0, t-\omega) \underline{\Lambda}'' d\omega P''(k-1, \omega) \quad (3.4.6)$$

The expression corresponding to the case where Q1 is empty can also be obtained by substituting $k=1$ in (3.4.6) as the busy period of Q2 starts with a single cell in that case. Removing the condition on the state of Q1 when the 1st cell arrives at an empty Q2 we get -

$$\frac{d}{dt} U_k''(t) = \frac{(1-p_0')}{D} \int_0^{Du(t-D_-)+tu(D-t)} P''(0, w) \underline{\Lambda}'' dw P''(k-1, t-w) + p_0' \delta_{1k} P''(0, t) \underline{\Lambda}'' \quad (3.4.7)$$

$$\underline{\Lambda}'' = I_M \otimes \Lambda'' \quad (3.4.8)$$

where I_M is the $M \times M$ identity matrix and p_0' is the probability that Q1 is empty at an arbitrary time instant. δ_{1k} denotes the Kronecker delta (1 if $k=1$ and 0 otherwise).

3.5. COMPUTATION OF $A_M'(t)$ AND $U_K'(t)$

Of these two quantities, the evaluation of $A_m'(t)$ is considered first. From (3.3.5) it can be observed that computation of $A_m'(t)$ requires the evaluation of $P'(m, t)$ and $dH'(t)$. $P'(m, t)$ can be evaluated using the recursive procedure given in Sec. 4.4. The expression for $\frac{d}{dt} H'(t)$ is obtained next.

Let us consider the inter departure time of cells from Q1 (IDT 1), i.e. $(\tau'_n - \tau'_{n-1})$, when Q1 is non-empty at the $(n-1)^{\text{th}}$ departure instant (i.e. $X'_{n-1} > 0$). It is the sum of a single cell service time for Q1 and the busy period of Q2 (BP) that might have been initiated by the cells arriving at Q2, if any,

in the interval $(\tau'_{n-1} - D, \tau'_{n-1})$. The number of cells arriving at Q2 in this interval depends on the phase of the MMPP $\underline{2}$ at $\tau'_{n-1} - D$. Hence IDT 1 depends on the phase of MMPP $\underline{2}$. Further IDT 1 is an integral multiple of D sec and an IDT 1 of t sec implies that BP of Q2 which follows a Q1 cell service time is equal to $(t-D)$ sec for t greater than D . Hence for evaluating $H'(t)$, an expression for the busy period distribution of Q2 is required and is obtained next.

For the computation of the busy period distribution of Q2 we define $MN \times MN$ matrices $\underline{G}^{(m)}(t)$ as follows, their $(i,j)^{th}$ elements denote the probability that the BP of Q2 which starts with m cells at time 0 with phase $\underline{j}(0)$ equal to i is of duration less than or equal to t sec and the phase of MMPP $\underline{2}$ at time t , i.e. $\underline{j}(t)$, is j . (We recall here that the phase of the composite phase process is equal to those of MMPP $\underline{1}$ and MMPP $\underline{2}$ at all time t). Let $\frac{d}{dt}\underline{G}^{(m)}(t)$, i.e. the probability density function of $\underline{G}^{(m)}(t)$, be denoted as $\underline{G}_k^{(m)}$ for $t=kD$, $k=1,2$, and let $\underline{G}_0^{(0)}$ to be the $MN \times MN$ identity matrix. Then $\underline{G}^{(m)}$ can be expressed as

$$\underline{G}^{(m)}(t) = \sum_{k=m}^{\infty} \underline{G}_k^{(m)} u(t-kD) \quad (3.5.1)$$

Here the lower limit for k is m because the BP which starts with m cells in Q2 will at least be of length mD sec. In Sec. 4.2, we develop a recursive procedure for the computation of $\underline{G}_k^{(m)}$.

We recall that the i^{th} diagonal element of the diagonal matrix $H'(t)$ denotes the probability that IDT 1 is of length less than or equal to t sec (or equivalently BP of Q2 is of length less than or equal to $t-D$ sec) given that the BP of Q2 starts at time 0 with phase $\underline{j}(0)$ of i . The pdf corresponding to the cumulative probability $H'(t)$ for $t = kD$, $k=1,2$ can be

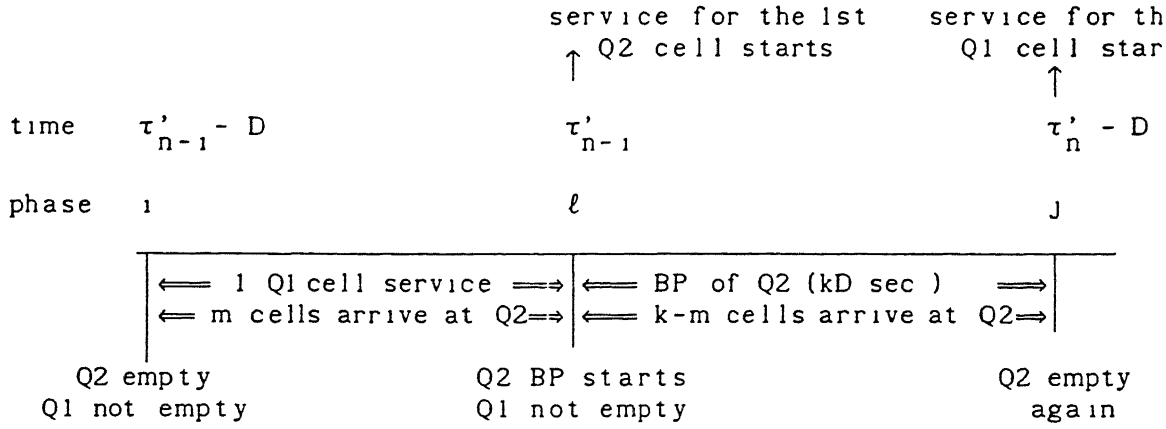


Fig 3.3 Arrivals and phase transitions at Q2 between adjacent Q1 service beginning epochs given Q1 not empty at the previous departure instant

obtained by using Fig 3.3 and considering the following chain of conditional events -

- (i) The phase of the MMPP \underline{z} at $\tau'_{n-1} - D$ is equal to 1 (i.e. $\underline{J}(\tau'_{n-1} - D)$ is 1) given that Q2 is empty
- (ii) Busy period of Q2 starts at time τ'_{n-1} with m cells in Q2 and with arrival phase equal to ℓ given that MMPP \underline{z} is in phase 1 at $\tau'_{n-1} - D$ (i.e. at the beginning of the service for the Q1 cell)
- (iii) The busy period is of duration kD sec and ends in phase j given that the busy period started with m cells and with arrival phase equal to ℓ

Let $c(k, 1)$, the 1^{th} element of the $MN \times 1$ vectors denote the joint probability that the BP of Q2 is of length kD sec and starts when MMPP \underline{z} is in phase 1

Then, using the above conditional events, $c(k, 1)$ is given by -

$$c(k, 1) = \sum_{m=(1-\delta_{0k})}^k \sum_{j=1}^{MN} \left(\frac{1}{p_0''} y_{0j}'' \right) \left(P_{j1}''(m, D) \right) \left(\sum_{\ell=1}^{MN} G_{k1\ell}''(m) \right) \quad (3.5.2)$$

where the 1^{th} element of the $MN \times 1$ vector y_0'' gives the probability of finding Q2 empty and MMPP \underline{z} in phase 1 at an arbitrary time instant and $p_0'' = y_0'' e$, where e is the $MN \times 1$ unit vector

The terms inside the first, second and third brackets of (3.5.2) denote respectively, the probability of occurrence of the events (i), (ii) and (iii) referred to above. The upper limit for m is k because any busy period which starts with more than k cells in Q_2 will be of duration greater than kD sec. The lower limit is 1 if there is indeed a busy period at all (i.e. k greater than 0) and is 0 otherwise.

Finally, Let $C(k)$ denote the $MN \times MN$ diagonal matrices whose i^{th} diagonal is equal to $P[BP = kD \text{ sec} \mid J'_{n-1} = i]$, i.e. the probability that BP of Q_2 is of length kD sec given that it started at time τ'_{n-1} with MMPP $\underline{2}$ in phase i . Using (3.5.2), it can be verified that

$$\begin{aligned} [C(k)]_{ij} &= P[BP = kD \mid J'_{n-1} = i] \delta_{ij} \\ &= c(k, i) \left[\sum_{n=1}^{\infty} c(n, i) \right]^{-1} \delta_{ij} \end{aligned} \quad (3.5.3)$$

As mentioned earlier an IDT 1 of t sec implies that the BP of Q_2 following the service of a Q_1 cell, is of length $t-D$ sec. Hence $\frac{dH'(t)}{dt}$ is given by

$$\frac{dH'(t)}{dt} = C(k-1) \delta'(t-kD) \quad (3.5.4)$$

Using (3.5.4) in (3.3.5) we get

$$A'_n(t) = \sum_{k=1}^{\infty} u(t-kD) C(k-1) P'(n, kD) \quad (3.5.5)$$

An expression for $\frac{d}{dt} U'_k(t)$ is derived next. As noted in Sec. 3.3, if Q_2 is not empty (n.e.) when the first cell arrives at an empty Q_1 , it waits, first, for the completion of the residual service time of the on going service, it then waits for the additional busy period (ABP) that the cells in Q_2 might require after the last service to be over before receiving service.

Let the residual service time and the ABP of Q_2 be denoted as w and uD ($i \geq 0$), respectively. Let the time at which Q_1 becomes idle, the time of the first cell arrival at an empty Q_1 , the time at which the ABP of Q_2 starts and the

time at which the busy period of Q1 starts be 0, $t-\epsilon D-\omega$, $t-\epsilon D$ and t , respectively, as shown in Fig 3.4. The arrival phase of MMPP 1 at these instants are i , ℓ , m and j respectively. Let the number of Q1 cells arriving during the residual service time of the present cell and the ABP be n and $k-n-1$, respectively. Then -

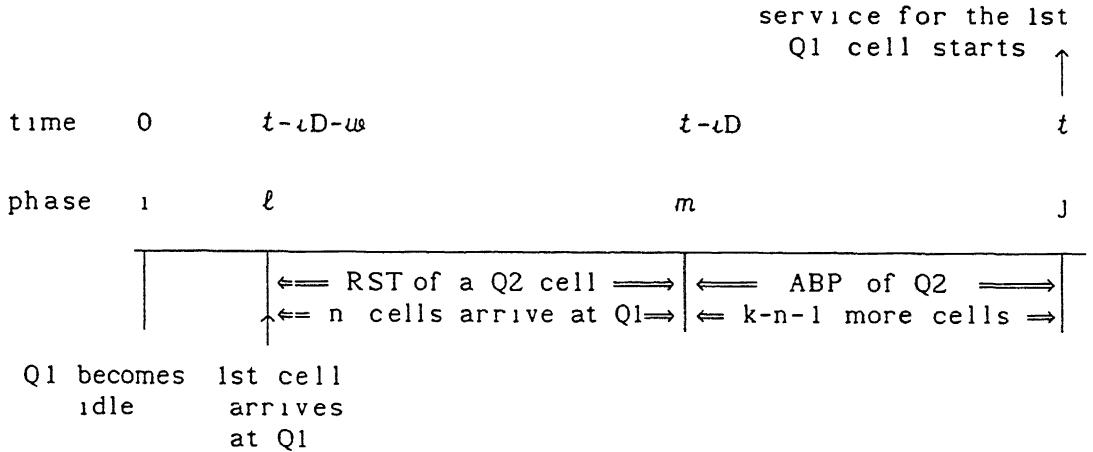


Fig 3.4 Time of occurrence of various events in the idle period of Q1 and the phase of the MMPP to Q1 at these instants

$$\left. \frac{d}{dt} U'_{k_{ij}}(t) \right|_{\substack{\text{RST}=\omega \\ \text{ABP}=\epsilon D \\ t=t, \text{ Q2 n e}}} = \sum_{n=n_0}^{k-1} \sum_{\ell=1}^{MN} \sum_{m=1}^{MN} P'_{i\ell}(0, t-\epsilon D-\omega) \Lambda'_{\ell\ell} d\omega P'_{\ell m}(n, \omega) P'_{mj}(k-n-1, \epsilon D) \quad (3.5.6)$$

Here, n_0 denotes the minimum value of n . The value of n_0 depends on the existence of the ABP following the Residual service time of a Q2 cell seen by the first cell arriving at an empty Q1. When ABP is not the present (i.e. $\epsilon=0$), then out of the k cells present at the time when BP of Q1 starts, $k-1$ cells arrive during the RST of the Q2 cell undergoing service. Hence the minimum value of n is equal to $k-1$ in this case. When ABP is non-zero, n_0 is zero and hence n_0 is given by

$$n_0 = (k-1)\delta_{\epsilon 0} \quad (3.5.7)$$

We also define the $MN \times MN$ diagonal matrices $F(k)$ whose i th diagonal element denotes $P[ABP = kD \text{ sec} | ABP \text{ starts at time } 0 \text{ with MMPP } \underline{2} \text{ in phase } i]$, i.e. the probability that an ABP of Q2 is of length kD sec given that it started at time 0 with MMPP $\underline{2}$ in phase i . Removing the condition on the length of ABP using the matrices $F(k)$, (3.5.6) becomes -

$$\left. \frac{d}{dt} U'_{k,ij}(t) \right|_{\substack{RST=\omega \\ t=t, \text{ Q2 n e}}} = \sum_{n=n_0}^{k-1} \sum_{\iota=0}^{\infty} \sum_{\ell=1}^{MN} \sum_{m=1}^{MN} P'_{\ell}(0, t-\iota D-\omega) \Lambda'_{\ell\ell} d\omega P'_{\ell m}(n, \omega) [F(\iota)]_{mm} P'_{m,j}(k-n-1, D) \quad (3.5.8)$$

This can be written in matrix form as

$$\left. \frac{d}{dt} U'_k(t) \right|_{\substack{RST=\omega, \\ t=t, \text{ Q2 n e}}} = \sum_{\iota=0}^{\infty} \sum_{n=n_0}^{k-1} P'(0, t-\iota D-\omega) \Lambda' d\omega P'(n, \omega) F(\iota) P'(k-n-1, \iota D) u(t-\iota D) \quad (3.5.9)$$

The residual service time of a Q2 cell seen by the first cell arriving at an empty Q1 is also uniformly distributed between the interval $(0, D)$ and hence its p.d.f. is given by (3.4.4). Removing the condition on the RST using (3.4.4) and noting that when $0 < t-\iota D < D$, the maximum RST is $t-\iota D$, we get -

$$\left. \frac{d}{dt} U'_k(t) \right|_{\substack{\text{Q2 n e} \\ t=t}} = \frac{1}{D} \sum_{n=n_0}^{k-1} \sum_{\iota=0}^{\infty} \int_0^{Du(t-\iota+1D_-) + (t-\iota D)u(\overline{\iota+1D}-t)} P'(0, t-\iota D-\omega) \Lambda'_{\ell\ell} d\omega P'(n, \omega) F(\iota) P'(k-n-1, \iota D) \quad (3.5.10)$$

We now consider the case where Q2 is empty when the first cell arrives at an empty Q1. In this case, the BP of Q1 always starts with a single cell. The corresponding expression for $U'_k(t)$ can be obtained using (3.5.10) by setting $k=1, \iota=0$. Combining the expressions corresponding to both these cases, we get

$$\frac{dU'_k(t)}{dt} = \sum_{\iota=0}^{\infty} \sum_{n=(k-1)\delta_{\iota 0}}^{k-1} \begin{cases} Du(t-\iota+1D_-) + (t-\iota D)u(\overline{\iota+1D}-t) \\ u(t-\iota D) \int_0^{Du(t-\iota+1D_-) + (t-\iota D)u(\overline{\iota+1D}-t)} P'(0, t-\iota D-\omega) \Lambda'_{\ell\ell} d\omega P'(n, \omega) \end{cases}$$

$$F(\iota) P'(k-n-1, \iota D) \left\} \frac{(1-p_0'')}{D} + p_0'' \delta_{1k} P'(0, t) \Lambda_1' \right. \quad (3.5.11)$$

In order to use (3.5.11), an expression for $F(\iota)$ needs to be derived. Analogous to the computation of the conditional probability distribution of the Busy Period (BP) viz $C(k)$, we define the $MN \times 1$ vectors $f(\iota)$ whose i^{th} element denotes the joint probability that the ABP is of duration ιD , and starts in phase i . Let the i^{th} element of the $MN \times 1$ vectors x_ι denote the steady state joint probability of finding ι cells in Q2 and the process MMPP \underline{z} in phase i at a departure instant of Q2. Note that if the ABP is to be of duration ιD sec, then the number of cells left in Q2 when the ABP starts can at most be ι . Using this and the definition of the matrices $G_\iota^{(m)}$ it can be verified that the i^{th} element of $f(\iota)$ and the $(i, j)^{\text{th}}$ element of $F(\iota)$ are given by -

$$[F(\iota)]_{i,j} = f(\iota, i) \left[\sum_{n=0}^{\infty} f(n, i) \right]^{-1} \delta_{ij} \quad (3.5.12)$$

$$f(\iota, i) = \sum_{m=1}^{\iota} \sum_{\ell=1}^{MN} x_{m1}'' G_{i, \ell}^{(m)} \quad (3.5.13)$$

3.6 TRANSFORMS OF THE ELEMENTS OF $Q'(\cdot)$ AND $Q''(\cdot)$

It is well known that the moments of a function can be computed by differentiating the appropriate transform of this function. In many cases, the computational effort required in this approach turns out to be small compared to the one where the function is differentiated directly. Hence we use the transform domain approach for the study of some of the characteristics of the busy periods of Q1 and Q2 as well as the moments of their queue lengths. For this, we need to evaluate the transforms of the elements of $Q'(\cdot)$ and $Q''(\cdot)$.

We start with the z-transform of $P'(m,t)$ and $P''(m,t)$ and state some of their properties. It can be shown using the results of section 5.4 of Neuts [3] that the z-transform $P''(m,t)$, denoted by $\mathcal{P}''(z,t)$, is given by -

$$\mathcal{P}''(z,t) = \sum_{n=0}^{\infty} P''(n,t)z^n = \exp\{R''(z)t\} \quad (3.6.1)$$

where $R''(z) = \underline{\Lambda}''z - \underline{\Lambda}'' + \underline{Q}^*$

Substituting $z=0$ and $z=1$ in (3.6.1), we get -

$$P''(0,t) = \exp\{(\underline{Q}^* - \underline{\Lambda}'')t\} \quad (3.6.2)$$

$$\sum_{n=0}^{\infty} P''(n,t) = \exp\{\underline{Q}^*t\} \quad (3.6.3)$$

Further, if the MMPP \underline{Q} is in phase i at time 0, it has to be in one of the phases j ($1 \leq j \leq MN$) at time t . Hence, the row sums of LHS of (3.6.3) are equal to 1 and hence $\exp\{\underline{Q}^*t\}$ is stochastic, i.e.

$$\exp\{\underline{Q}^*t\} \mathbf{e} = \mathbf{e} \quad (3.6.4)$$

The corresponding expressions for the z-transform of $P'(n,t)$, denoted as $\mathcal{P}'(z,t)$, and $P'(0,t)$ can be obtained using (3.6.1) and (3.6.2) by replacing the parameters of Q_2 by those of Q_1 .

Next, the LST of $A_m''(t)$ and $U_k''(t)$, denoted as $\tilde{A}_m''(s)$ and $\tilde{U}_k''(s)$ respectively, are obtained. The LST of a function is denoted by that function with a superscript of tilde (\sim). Using (3.4.2), $\tilde{A}_m''(s)$ is obtained as-

$$\tilde{A}_m''(s) = P''(m,D)\exp\{-sD\} \quad (3.6.5)$$

$\tilde{U}_k''(s)$ is evaluated using (3.4.7) and (3.6.2). Considering the case where Q_1 is non-empty (n.e.) when the 1st cell arrives at an empty Q_2 we get -

$$\tilde{U}_k''(s) \Big|_{\substack{Q_1 \\ n.e.}} = \frac{1}{D} \int_0^{\infty} dt e^{-st} \int_0^t e^{(\underline{Q}^* - \underline{\Lambda}'')(t-w)} \underline{\Lambda}'' dw P''(k-1,w) \quad (3.6.6)$$

$$= \frac{1}{D} \int_0^D dt e^{-t(sI - R''(0))} \int_0^t dw T''(w,k) + \frac{1}{D} \int_D^{\infty} dt e^{-t(sI - R''(0))} \int_0^D dw T''(w,k) \quad (3.6.7)$$

Proceeding along the same lines as for the computation of $\tilde{U}_k''(s)$, $\tilde{U}_k'(s)$ can be computed using (3.5.11) and is given by -

$$\begin{aligned} \tilde{U}_k'(s) = & [sI - R'(0)]^{-1} (1 - p_0'') \frac{1}{D} \sum_{n=0}^{k-1} \sum_{\iota=0}^{\infty} \int_0^D \Lambda' P'(n, w) dw e^{-s(w + \iota D)} F(\iota) P'(k-n-1, \iota D) \\ & + [sI - R'(0)]^{-1} p_0'' \frac{\Lambda'}{D} \delta_{1k} \end{aligned} \quad (3.6.14)$$

Using (3.3.18) and (3.6.14), $\tilde{B}_m'(s)$ the LST of $B_m'(t)$ can be obtained as -

$$\tilde{B}_m'(s) = \sum_{k=1}^{m+1} \tilde{U}_k'(s) e^{-sD} P'(m-k+1, D) \quad (3.6.15)$$

One immediate application of these transforms would be the computation of $Q'(\omega)$ and $Q''(\omega)$. Substituting $s=0$ in (3.6.5), (3.6.11), (3.6.13) and (3.6.14), $\tilde{A}_m''(\omega)$, $\tilde{U}_k''(\omega)$, $\tilde{A}_m'(\omega)$ and $\tilde{U}_k'(\omega)$ can be computed. These matrices are required for the computation of the QLDs as discussed in Sec. 5.6. Using these transforms, we can also compute the double transforms (z transform of the LST) of $A_m''(t)$, $U_k''(t)$, $A_m'(t)$, $U_k'(t)$, $B_m''(t)$ and $B_m'(t)$ denoted as $\tilde{A}''(z, s)$, $\tilde{U}''(z, s)$, $\tilde{A}'(z, s)$, $\tilde{U}'(z, s)$, $\tilde{B}''(z, s)$, and $\tilde{B}'(z, s)$ respectively. In the subsequent sections, we find these transforms useful for examining the stability of the queues Q1 and Q2 and for studying their busy period characteristics. These are also required for the computation of the probabilities p_0'' and p_0' of the queues Q2 and Q1 being empty.

Using (3.6.1) and (3.6.5) it can be verified that -

$$\tilde{A}''(z, s) = \sum_{m=0}^{\infty} \tilde{A}_m''(s) z^m = e^{[R''(z) - sI]D} \quad (3.6.16)$$

Similarly using (3.6.13) and (3.6.1) we get -

$$\tilde{A}'(z, s) = \sum_{m=0}^{\infty} \tilde{A}_m'(s) z^m = \sum_{k=1}^{\infty} C(k-1) e^{[R'(z) - sI]kD} \quad (3.6.17)$$

For the computation of $\tilde{U}''(z, s)$, we consider the case where the first cell

arriving at an empty Q2 finds Q1 not empty Using (3.6.11) and expanding the summation wrt n we get -

$$\tilde{U}''(z,s) \Big|_{Q1 \text{ n e}} = \sum_{m=0}^{\infty} \tilde{U}_m''(s) z^m \Big|_{Q1 \text{ n e}} \quad (3.6.18)$$

$$= [sI - R''(0)]^{-1} \frac{1}{D} \int_{w=0}^D e^{-sw} \underline{\Lambda}'' dw [P''(0,w)z + P''(1,w)z^2 + \dots]$$

$$= [sI - R''(0)]^{-1} \frac{1}{D} \int_{w=0}^D \underline{\Lambda}'' z dw \mathcal{P}''(z,w) e^{-sw} \quad (3.6.19)$$

Considering the case where Q1 is empty when the first cell arrives at Q1 and removing the conditioning on the state of Q1 we obtain -

$$\tilde{U}''(z,s) = [sI - R''(0)]^{-1} \left\{ \underline{\Lambda}'' z p'_0 + (1-p'_0) \frac{1}{D} \int_{w=0}^D \underline{\Lambda}'' z \mathcal{P}''(z,w) e^{-sw} dw \right\} \quad (3.6.20)$$

The evaluation of $\tilde{U}'(z,s)$ is carried out along the same lines as for the computation of $\tilde{U}''(z,s)$ Using (3.6.14), expanding the summations on the variables n and ι and combining the like terms, we get -

$$\tilde{U}'(z,s) \Big|_{Q2 \text{ n e}} = \sum_{m=0}^{\infty} \tilde{U}_m'(s) z^m \Big|_{Q2 \text{ n e}} \quad (3.6.21)$$

$$= [sI - R'(0)]^{-1} \frac{1}{D} \sum_{\iota=0}^{\infty} \int_{w=0}^D e^{-s(w+\iota D)} \underline{\Lambda}' dw \left\{ P'(0,w)F(\iota)P'(0,\iota D)z + \{P'(0,w)F(\iota)P'(1,\iota D) + P'(1,w)F(\iota)P'(0,\iota D)\}z^2 + \dots \right\} \quad (3.6.22)$$

$$= [sI - R'(0)]^{-1} \frac{1}{D} \sum_{\iota=0}^{\infty} \int_{w=0}^D \underline{\Lambda}' z dw e^{-s(w+\iota D)} \mathcal{P}'(z,w)F(\iota)\mathcal{P}'(z,\iota D) \quad (3.6.23)$$

Considering the case where Q2 is empty when the first cell arrives at Q1 and removing the conditioning on the state of Q2 we obtain -

$$\begin{aligned}\tilde{U}'(z,s) &= [sI - R'(0)]^{-1}(1-p_0'') \frac{1}{D} \sum_{\iota=0}^{\infty} \int_{w=0}^D \underline{\Lambda}'z \, dw e^{-s(w+\iota D)} \mathcal{P}'(z,w)F(\iota)\mathcal{P}'(z,\iota D) \\ &\quad + [sI - R'(0)]^{-1} \underline{\Lambda}'z p_0''\end{aligned}\quad (3.6.24)$$

Using (3.6.12) and (3.6.15) and combining the like terms as in (3.6.22) we get

$$\tilde{B}''(z,s) = \tilde{U}''(z,s)\mathcal{P}''(z,D)e^{-sD} \quad (3.6.25)$$

$$\tilde{B}'(z,s) = \tilde{U}'(z,s)\mathcal{P}'(z,D)e^{-sD} \quad (3.6.26)$$

Finally, we consider the evaluation of these transforms when $z=1$ and $s=0$

Using theorem (5.3.2) of Neuts[3], it can be verified that -

$$[\underline{\Lambda}' - \underline{Q}^*]^{-1} \underline{\Lambda}'e = [\underline{\Lambda}'' - \underline{Q}^*]^{-1} \underline{\Lambda}''e = e \quad (3.6.27)$$

Using (3.6.27) and (3.6.4), it can be shown that $\tilde{A}''(1,0)$, $\tilde{A}'(1,0)$, $\tilde{U}''(1,0)$ and $\tilde{U}'(1,0)$ are stochastic matrices

REFERENCES.

- 1 L Kleinrock, "Queueing systems, Vol 1 Theory ", Wiley, New York, 1975
- 2 M F Neuts, " Matrix-Geometric Solutions in Stochastic models An Algorithmic approach", Johns Hopkins Univ Press, Baltimore, 1981
- 3 M F Neuts, "Structured stochastic matrices of the M/G/1 type and their applications, Marcel Dekker, New York, 1989
- 4 R Bellman, "Introduction to Matrix Analysis", McGraw-Hill, New York, 1960
- 5 V Ramaswami, "The N/G/1 queue and its detailed analysis", Adv Appl Prob 12 (1980), pp 222-261
- 6 M F Neuts, "A versatile Markovian point process, J Appl Prob 16 (1979) pp 764-779

CHAPTER 4

CHARACTERISTICS OF BUSY PERIODS OF Q1 AND Q2

4.1. INTRODUCTION

In the study of queueing systems with arrivals from more complex processes like Neuts process, MMPP etc, the so called $G()$ matrices and their moment matrices play a crucial role (eg Neuts [1], Ramaswami [2], Lucantoni[3]). Their role is analogous to that of the busy period and its transform in the study of simple M/G/1 systems. In Ramaswami [2], these matrices were used to study some of the characteristics of N/G/1 queue. In this chapter, we define appropriate $G()$ matrices for studying some of the characteristics of the queues Q1 and Q2 (as described in the earlier chapters) and consider the application of the results of Neuts[1] and Ramaswami[2] to the problem considered in this thesis.

4.2 FIRST PASSAGE TIMES AND BUSY PERIODS OF $Q'()$ AND $Q''()$

First, we motivate the need for the study of the busy period characteristics for the study of the QLD of a queue. Let us consider the evaluation of the QLD of the simple M/G/1 queue using embedded Markov chain approach. Let us choose the embedded points to be the departure instants of customers from this queue. Let X_n denote the number of customers in the system at the n^{th} departure instant from the queue. The sequence $\{X_n, n \geq 0\}$ forms a semi-Markov chain (SMC) and let the transition probability of this SMC be denoted as Q with $(i,j)^{\text{th}}$ elements as Q_{ij} . Let the steady state probability of finding i customers at a departure instant be denoted as x_i . These system probabilities can be evaluated by solving the system of equations given by-

$$x_j = \sum_{i=0}^{\infty} x_i Q_{ij}, \quad \sum_{j=0}^{\infty} x_j = 1 \quad (j = 0, 1, 2, \dots) \quad (4.2.1)$$

There is an alternate method for finding the x_j 's (See for e.g. Takacs[4], Feller [5]) We shall consider this method next. We shall assume that the traffic arrival rate to the system is less than the service rate and hence the system is stable. Under this condition, it can be noted that the above Markov chain is irreducible as starting from any state ($X_n = j$) any other state ($X_m = i$) can be reached for $i \geq 0$ and $m > n$. It is also aperiodic. Let us assume that the system was initially in the state $E_1 = \{X_n = i\}$. Let $f_1^{(m)}$ denote the probability that the system returns to state E_1 again after m transitions (i.e. after m departures). The mean recurrence time of the state E_1 is denoted as μ_1 and is given by

$$\mu_1 = \sum_{m=1}^{\infty} m f_1^{(m)} \quad (4.2.2)$$

As the system is stable, $\mu_1 < \infty$ and E_1 is a non-null state. As the system is stable, E_1 is also a recurrent state i.e.

$$f_1 = \sum_{m=1}^{\infty} f_1^{(m)} = 1 \quad (4.2.3)$$

In an irreducible and aperiodic Markov chain with recurrent null states x_j is given by (see for e.g. [4], [5])

$$x_j = \frac{1}{\mu_j} \quad (4.2.4)$$

Hence, in principle, by finding the mean recurrence times of the various states, the system probabilities can be found. The need for finding at least some of the system probabilities using this method may arise due to the following reason. When the traffic offered to the queue is close to the capacity of the server, the dimension of Q may become very large and hence the system of equations given by (4.2.1) may have to be solved recursively. The recursive

procedures require the knowledge of the probabilities x_0 and x_1 to start with

Let us next illustrate how the mean recurrence times μ_i for $i = 0, 1$, can be found. Let us consider the case when $i=0$ first. i.e. The system starts in the state $E_0 = \{X_n = 0\}$. This state implies that the system has become empty. For this state to be visited again at least one customer has to arrive at the system and the busy period has to start first. When this busy period ends the state E_0 is visited again. Hence the recurrence time of this state is equal to the number of customers served during the busy period. The mean recurrence time is obviously equal to the average number of customers served during a busy period. Hence the characteristics of the busy period is required for computing x_0 .

To illustrate the computational complexity associated with the evaluation of the higher system probabilities using (4.2.4), evaluation of the MRT of E_1 is considered next. For computing the MRT of E_1 , the following events have to be considered. To start with, the system state is E_1 . This state is visited again if one of the three things happen

- (1) During the service time that follows, exactly one customer arrives
- (2) At the next departure instant the system became empty. In one of the subsequent busy cycles (idle period followed by busy period), the state E_1 is visited again
- (3) During the service that follows, more than one customer arrives. In this case the state E_1 is visited again before the ongoing busy period ends and the recurrence time is greater than 1

We shall not proceed with the evaluation of the MRT of E_1 . Next we list the events that need to be considered for the evaluation of the MRT of E_2 . To start with the system state was E_2 . This state is visited again when one of the following three set of events occur

- (1) exactly one customer arrived at the system in the service time that follows
- (2) no customer arrived in the service time that follows and one of the following two events occur
 - (i) The state E_2 is visited again at one of the subsequent departure instants before the ongoing busy period ends
 - (ii) The state E_2 is not visited again in the ongoing busy period and it is visited again only in one of the subsequent busy cycles
- (3) More than one customer arrived at the system in the service time that follows, during a later departure instant in the ongoing busy period, the state E_3 is visited again

From these list of events, it can be observed that the number of events that has to be taken into account for the computation of the MRT of E_1 increases as the value of λ increases. Hence the computation of the system probabilities using (4.2.4) is not preferred for large values of λ .

Having noted the role played by the characteristics of the busy period in an M/G/1 queue, next we consider the generalizations required for the study of prioritized queueing system. Let us concentrate on Q2 first.

- (1) The system probabilities depend on the phase of the arrival process and hence x_i 's have to be considered as vectors whose i^{th} element denote the joint probability that $X_n'' = i$ and the phase of the arrival process $= j$. We shall define the system to be in level $i = \{1, j, 1 \leq j \leq MN\}$ when the number of customers in the system $= i$ and the phase of the arrival process $= j$.
- (2) For brevity of notation we denote the semi-Markov Process whose transition probability matrix is $Q''()$ as the semi-Markov process $Q''()$. We define the time taken for the semi-Markov process $Q''()$ to go from the set of states $\{i+1\} = \{i+1, j, 1 \leq j \leq MN\}$ to the set of states $\{i\} = \{i, j, 1 \leq j \leq MN\}$ as the first passage time.

(3) The busy period can start with more than one customer in the system. The busy period distribution depends on the phase of the arrival process at the beginning of the busy period and hence should be treated as matrix functions. We assume the busy period of Q2 to be characterized by $MN \times MN$ matrices $G^{(1)}(k, t)$. Their (j, j') th element denotes the probability that, given that the semi-Markov process $Q^{(1)}$ starts in the state $(1, j)$, it reaches the state $(0, j')$ for the first time after k transitions and the time taken for such a first passage is at most t .

(4) We denote $G^{(1)}(k, t)$ as $G^{(1)}(k, t)$. These matrices for $k=0, 1, 2, \dots$ specify completely the first passage time distribution of $Q^{(1)}$, i.e. the time taken for $Q^{(1)}$ to return to level 0 starting from level 1. (Here, the time is expressed in terms of the number of customer services).

From our discussion on the computation of the MRTs for the M/G/1, it is clear that for the prioritized queue as well, we need to compute the statistics like the number of customers served during the busy period. In section (3.5), we noted that the busy period distribution of the higher priority queue is required to compute the distribution of the inter-departure time of customers from Q1. Efficient procedures for the evaluation of these statistics can be arrived at using the double transform (the z transform of the LST) of $G^{(1)}(k, t)$ denoted as $\tilde{G}^{(1)}(z, s)$, i.e. $\tilde{G}^{(1)}(z, s)$ is given by-

$$\tilde{G}^{(1)}(z, s) = \sum_{k=0}^{\infty} z^k \int_0^{\infty} dG^{(1)}(k, t) \exp(-st) \quad (4.2.5)$$

The double transform of $G^{(1)}(k, t)$ is denoted as $\tilde{G}^{(1)}(z, s)$. The matrices $G^{(1)}(k, t)$ are similar to the matrices $G(k, x)$ defined in Neuts [1,3] and Ramaswami [2] and hence $\tilde{G}^{(1)}(z, s)$ satisfy the following properties

(i) In view of the structure of $Q^{(1)}$, the first passage times from the set of states $\{i+1\}$ to $\{i\}$ are identically distributed and hence -

$$\tilde{G}^{(1)}(z,s) = [\tilde{G}''(z,s)]^1 \quad (4.2.6)$$

(ii) $\tilde{G}''(z,s)$ satisfies the non-linear matrix equation

$$\tilde{G}''(z,s) = z \sum_{m=0}^{\infty} \tilde{A}_m''(s) \tilde{G}^{(m)}(z,s) \quad (4.2.7)$$

where $\tilde{A}_m''(s)$ is the LST of $A_m''(t)$

(iii) For $0 \leq z \leq 1$, $s > 0$, there exists a unique non-negative matrix $\tilde{G}''(z,s)$ which satisfies (ii). This solution is such that $G'' = G''(1,0)$ is an irreducible sub-stochastic matrix. (The Markov Renewal process $Q''(\cdot)$ is said to be irreducible iff there is a positive probability of visiting the state $(i-1, j')$, $1 \leq j, j' \leq MN$ from any initial state (i, j)). The matrix G'' is the entry wise smallest non-negative solution of the nonlinear matrix equation

$$G'' = \sum_{n=0}^{\infty} \tilde{A}_n''(0) G''^n \quad (4.2.8)$$

(iv) Let ρ'' denote the average traffic offered at Q2. Then $Q''(\cdot)$ is recurrent, G'' is stochastic and Q2 is stable if $\rho'' < 1$. Analogous to that in an M/G/1 queue, ρ'' can be obtained as

$$\rho'' = \pi'' \beta'' \quad (4.2.9)$$

where π'' is the $1 \times MN$, invariant probability vector of $\tilde{A}''(1,0)$. The $(i, j)^{th}$ element of $\tilde{A}''(1,0)$ gives the probability of finding the MMPP \underline{z} in phase j at a departure instant of Q2 given that the previous departure from Q2 left the system non-empty and MMPP \underline{z} in phase i . The i^{th} element of the $MN \times 1$ vector, β'' denotes the average number of customers arriving at Q2 during the inter-departure time of customers from Q2 given that at the previous departure instant Q2 is non-empty and the MMPP \underline{z} is in phase i .

Using these properties computation of the busy period distribution of Q2 is considered in section (4.3). The statistics on the number of customers served during a busy period is considered in section (4.5). The matrix G'' plays a central role in the computation of the first passage times. The

$(i,j)^{\text{th}}$ element of G'' denotes the probability that the busy period of Q2 ends in phase j given that the busy period started with the MMPP \underline{Q} in phase i

Next we consider the computation of G'' . For the case where the interdeparture time distribution of the customers from the queue (IDT) is a scalar function, an iterative procedure for the computation of G is given in Lucantoni [6]. This iterative procedure has been generalised for the present case where the IDT is defined to be a matrix function. Using (3.4.1), the k^{th} iterate of G'' , denoted as G_k'' can be obtained using the following equations -

$$G_0'' = 0 \quad G_{k+1}'' = \sum_{n=0}^{\infty} \Gamma_n'' H_{n,k}'' \quad (4.2.10)$$

$$H_{0,k}'' = I \quad H_{n+1,k}'' = [I - \theta''^{-1} R''(0)] H_{n,k}'' + \theta''^{-1} \Lambda'' H_{n,k}'' G_k'' \quad (4.2.11)$$

$$\theta'' = \max(\Lambda_{jj}'' - Q_{jj}^*) \quad 1 \leq j \leq MN \quad (4.2.12)$$

$$\Gamma_n'' = \int_0^{\infty} e^{-\theta'' t} \frac{(\theta'' t)^n}{n!} dH''(t) \quad (4.2.13)$$

$$= e^{-\theta'' D} \frac{(\theta'' D)^n}{n!} I \quad (4.2.14)$$

Let $\Gamma_0'' = e^{-\theta'' D} I$, then Γ_n'' can be computed recursively as follows

$$\Gamma_n'' = \Gamma_{n-1}'' \frac{\theta'' D}{n} \quad (4.2.15)$$

Following an approach similar to that for Q2, we define for Q1 the $MN \times MN$ matrices $G^{(1)}(k,t)$ whose (j,j') th element denotes the probability that, given that the semi-Markov process $Q'()$ starts in the state (i,j) , it reaches the state $(0,j')$ for the first time after k transitions and the time of such a first passage is at most t . The results (i) - (iv) given earlier for Q2 are

also valid for Q1 when the parameters of Q2 are replaced by those of Q1. Similarly G' can also be computed using (4.2.10) - (4.2.13). Γ'_n can be obtained using (3.5.4) as follows

$$\Gamma'_n = \int_0^{\infty} e^{-\theta't} \frac{(\theta't)^n}{n!} dH'(t) \quad (4.2.16)$$

$$= \int_0^{\infty} \sum_{m=0}^{\infty} e^{-\theta't} \frac{(\theta't)^n}{n!} C(m) \delta(t-mD) dt \quad (4.2.17)$$

$$= \sum_{m=0}^{\infty} e^{-\theta'mD} \frac{(\theta'mD)^n}{n!} C(m) = \sum_{m=0}^{\infty} C(m) \Gamma'_{m,n} \quad (4.2.18)$$

Here, D denotes the service time/cell and $C(m)$ s are the matrices defined in section (3.5). Let $\Gamma'_{m,0} = e^{-\theta'mD}$, then $\Gamma'_{m,n}$ can be recursively computed for any particular value of m starting from $n=0$ as follows

$$\Gamma'_{m,n} = \Gamma'_{m,n-1} \frac{(\theta'mD)}{n} \quad (4.2.19)$$

4.3. THE BUSY PERIOD DISTRIBUTION OF Q1 AND Q2

As mentioned in Sec. 3.5, for computing the elements of $Q'()$, the busy period distribution of the server in Q2 must be known. The busy period distribution can be computed using Ramaswami [7] and this requires the computation of the inverse LST. However, taking advantage of the fact that the service time per cell is constant, an efficient procedure for the computation of the busy period distribution of Q2 is developed here which does not require the computation of inverse transforms. The busy period distribution of Q1 can also be computed along the same lines.

It can be noted that the LST of the busy period distribution of Q2 can be obtained using (4.2.6), (4.2.7) and (3.6.5) by substituting $z=1$ as follows,

$$\tilde{G}''(1,s) = \sum_{m=0}^{\infty} A_m''(s) \tilde{G}''^{(m)}(1,s) \quad (4.3.1)$$

$$= \sum_{m=0}^{\infty} P''(m,D) e^{-sD} \left[\tilde{G}''(1,s) \right]^m \quad (4.3.2)$$

It may be noted that the matrix $\underline{G}''^{(m)}(t)$, defined in section (3.5), for $m=1$ gives the distribution of the busy period of Q2 starting with a single customer. Hence the LST of $\underline{G}''^{(1)}(t)$ is also equal to $\tilde{G}''(1,s)$. Using (3.5.1), $\tilde{G}''(1,s)$ can be written as

$$\tilde{G}''(1,s) = \sum_{\ell=1}^{\infty} G_{\ell}''^{(1)} e^{-sD\ell} \quad (4.3.3)$$

Let $\tilde{G}''(1,s)$, e^{-sD} and $G_{\ell}''^{(1)}$ be denoted as $G(\xi)$, ξ and G_{ℓ}'' , respectively

Then, using (4.3.3) in (4.3.2), we get

$$\sum_{\ell=1}^{\infty} G_{\ell}'' \xi^{\ell} = \sum_{m=0}^{\infty} P''(m,D) \xi \left[\sum_{\ell=1}^{\infty} G_{\ell}'' \xi^{\ell} \right]^m \quad (4.3.4)$$

Because of the appearance of a lone " ξ " on the RHS of (4.3.4), by comparing the coefficients of ξ^n on both sides of (4.3.4), it can be verified that

$$G_1'' = P''(0,D) \quad (4.3.5)$$

$$G_n'' = \sum_{k=0}^{\infty} P''(k,D) \left\{ \text{coefficient of } \xi^{n-1} \text{ in } \left[\sum_{\ell=1}^{\infty} G_{\ell}'' \xi^{\ell} \right]^k \right\} \quad (4.3.6)$$

Let the coefficient of $(\xi^1 \xi^2 \xi^3 \xi^4 \xi^5 \dots)$ in $(G_1'' \xi + G_2'' \xi^2 + G_3'' \xi^3 + \dots)^k$ be represented as a matrix array $Y(k)$ whose i^{th} element $Y(k,i)$ gives the coefficient of ξ^i . It may be noted that the coefficient of ξ^n in $(G_1'' \xi + G_2'' \xi^2 + G_3'' \xi^3 + \dots)^{\ell}$ is zero for $n < \ell$ i.e.

$$Y(\ell,n) = 0 \text{ for } n < \ell \quad (4.3.7)$$

Using the array representation and using (4.3.7), (4.3.6) can be rewritten as

$$G_n'' = Y(1,n) = \sum_{m=0}^{n-1} P''(m,D) Y(m,n-1) \quad (4.3.8)$$

A recursive procedure for the computation of $P''(m,D)$ is developed in the next section. If these are known, our problem becomes one of computing the elements of the matrix arrays $Y(m)$'s. It may be noted that the coefficient

of ξ^n in $(G_1''\xi + G_2''\xi^2 + G_3''\xi^3 + \dots)^2$ can be obtained by convolving the sequence $(G_1'' G_2'' G_3'' \dots)$ with $(G_1'' G_2'' G_3'' \dots)$ and is given by

$$Y(2,n) = \sum_{k=1}^{n-1} G_n'' G_{n-k}'' = \sum_{k=1}^{n-1} Y(1,k) Y(1,n-k) \quad (4.3.9)$$

Similarly, the coefficient of ξ^n in $(G_1''\xi + G_2''\xi^2 + G_3''\xi^3 + \dots)^3$ can be obtained by convolving the sequence $(Y(2,1) Y(2,2) Y(2,3) \dots)$ with $(G_1'' G_2'' G_3'' \dots)$ and is given by

$$Y(3,n) = \sum_{k=1}^{n-1} Y(2,k) Y(1,n-k) \quad (4.3.10)$$

Generalising this result, the coefficient of ξ^n in $(G_1''\xi + G_2''\xi^2 + G_3''\xi^3 + \dots)^m$ can be obtained by convolving the sequence $(Y(m-1,1) Y(m-1,2) Y(m-1,3) \dots)$ with $(G_1'' G_2'' G_3'' \dots)$ and is given by

$$Y(m,n) = \sum_{k=1}^{n-1} Y(m-1,k) Y(1,n-k) \quad (4.3.11)$$

Finally let $Y(0,n)$ be defined as

$$Y(0,n) = I \delta_{n0} \quad (4.3.12)$$

We then compute $G_n'' = Y(1,n)$ for $n=1,2,3$ as follows -

(i) $Y(1,1)$ is given by (4.3.5). Using (4.3.7), we note that $Y(m,1) = 0$ for $m > 1$. Hence $Y(m,1)$ is known for $m=1,2,3$.

(ii) Knowing $Y(0,1)$ and $Y(1,1)$, we can then find $Y(m,2)$ for $m=1,2$, using (4.3.8), (4.3.12) and (4.3.7).

(iii) Knowing $Y(0,2)$, $Y(1,2)$ and $Y(2,2)$, we find $Y(m,3)$ for $m=1,2$, next using (4.3.8), (4.3.12) and (4.3.7).

(iv) Proceeding in this manner, at the n^{th} step we find $Y(m,n)$ for $m=1,2$, using (4.3.8), (4.3.12), (4.3.7) and the values of $Y(0,n-1)$, $Y(1,n-1)$ and $Y(2,n-1)$, $Y(n-1,n-1)$ computed in the $(n-1)$ th step.

Using this procedure, G_n'' for $n=1,2$, can be found recursively. Moreover, this recursive procedure also yields $Y(k)$ and hence $[\tilde{G}(1,s)]^k$. In

light of (4.2.6), $Y(k,n)$'s are nothing but the $G_n^{(k)}$'s. Hence, as a by-product, this recursive procedure also enables the computation of the matrices $C(k)$ and $F(k)$ defined in Sec. 3.5

We can now consider the application of these results for an MMPP/D/1 queue with a FCFS service discipline. Let the parameters of this queue be denoted by the corresponding parameters of Q2 without the superscript of ("). In this case, the busy period always starts with a single cell. $A_m(t)$ and $G^{(1)}(t)$ are of the same form as (3.4.2) and (3.5.1) respectively and hence the above recursive procedure is also valid for this queue. Let $G_n = G_n^{(1)}$. Then it can be shown that the probability that the busy period of this queue is equal to nD sec is given by

$$P(BP=nD) = \frac{y_0 G_n e}{y_0 e} \quad (4.3.13)$$

If the arrival parameters of this MMPP/D/1 FCFS queue are chosen to be equal to those of Q2, then G_n is equal to G_n'' . Hence the busy period distribution of Q2, given that the busy period starts with a single cell in Q2, is the same as that of an equivalent MMPP/D/1 queue with FCFS discipline.

Next, the extension of the above results to Q1 is considered. For this purpose, analogous to that for Q2, we define $MN \times MN$ matrices $G_n^{(k)}$ for $k, n = 1, 2, \dots$. The $(i,j)^{th}$ element of these matrices gives the probability that the busy period of Q1 is of duration nD sec and ends with MMPP $\underline{1}$ in phase j given it started at time 0 with k customers in Q1 and with MMPP $\underline{1}$ in phase i . For ease of notation let $G_n^{(1)}$ be denoted as G_n' . Using these matrices $\tilde{G}'(1,s)$ can be written as

$$\tilde{G}'(1,s) = \sum_{\ell=1}^{\infty} G_{\ell}'^{(1)} e^{-sD\ell} \quad (4.3.14)$$

By replacing the parameters of Q2 by those of Q1 in (4.2.7) it can be verified

that $\tilde{G}'(1,s)$ satisfies the equation

$$\tilde{G}'(1,s) = \sum_{m=0}^{\infty} \tilde{A}'_m(s) \tilde{G}'^{(m)}(1,s) \quad (4.3.15)$$

$$= \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} C(k-1)P'(m,kD)e^{-skD} \left[\tilde{G}'(1,s) \right]^m \quad (4.3.16)$$

(4.3.16) is obtained from (4.3.15) by substituting the expression for $\tilde{A}'_m(s)$ using (3.6.13). Let $\xi = e^{-sD}$. Substituting (4.3.14) in (4.3.16) we get

$$\sum_{\ell=1}^{\infty} G_{\ell}'^{(1)} \xi^{\ell} = \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} C(k-1)P'(m,kD) \xi^k \left[\sum_{\ell=1}^{\infty} G_{\ell}'^{(1)} \xi^{\ell} \right]^m \quad (4.3.17)$$

Comparing the coefficients of ξ^n on both sides of (4.3.17), G_n' is given by

$$G_n' = \sum_{k=1}^n \sum_{m=0}^{n-k} C(k-1)P'(m,kD) \left\{ \text{coeff of } \xi^{n-k} \text{ in } \left[\sum_{\ell=1}^{\infty} G_{\ell}' \xi^{\ell} \right]^m \right\} \quad (4.3.18)$$

Since for $k > n$, the exponent of ξ becomes less than 0, the upper limit on k is set to be n . Similarly, it can be shown that the upper limit of m is $n-k$. Analogous to the procedure outlined for Q_2 , we can recursively obtain G_n' for $n=1,2,3$, by defining similar matrix arrays

4.4. COMPUTATION OF $P''(M,D)$ AND $P'(M,ND)$

We consider the computation of the matrices $P''(m,D)$ first. One direct way of obtaining these matrices is to solve the differential-difference equations satisfied by the matrices $P''(m,t)$ (e.g. Neuts [1], [8]). However, these equations do not have closed form solutions and hence they have to be solved through numerical integration. Since the other parameters like A_m'' 's and U_k'' 's have to be evaluated using these matrices, these numerical integrations have to be performed with a very high degree of accuracy and require a lot of computational effort. We next show how these matrices can also be evaluated efficiently using the recursive procedure suggested for the computation of the

busy period distribution

We start with the expression for the z-transform of $P''(m,D)$ in terms of the parameters of the arrival process (MMPP 2) -

$$P''(z,D) = \sum_{m=0}^{\infty} P''(m,D) z^m = e^{[\Lambda(z-1)+Q^*]D} \quad (4.4.1)$$

This can be rewritten in terms of the infinite series expansion as

$$\sum_{m=0}^{\infty} P''(m,D) z^m = \sum_{m=0}^{\infty} \frac{1}{m!} \{(\Lambda''(z-1)+Q^*)D\}^m \quad (4.4.2)$$

It is now obvious that $P''(m,D)$ can be evaluated by comparing the coefficients of z^n on both sides of (4.4.2). Let the coefficient of (z^0, z^1, z^2, \dots) in $\{[\Lambda''(z-1)+Q^*]D\}^n$ be represented as a matrix array $V(n)$ whose i^{th} element $V(n,i)$ gives the coefficient of z^{i-1} . Let $V(0)$ be defined to the matrix array with $V(0,n) = I \delta_{0n}$. For $n \geq 2$, the elements of $V(n)$ can now be obtained by convolving $V(n-1)$ with $V(1)$ and hence $V(n,m)$ and $P''(m,D)$ are given by

$$V(n,m) = \sum_{k=1}^2 V(1,k) V(n-1, m-k+1) \quad (4.4.3)$$

$$V(n,m) = 0 \text{ for } m > n+1 \quad (4.4.4)$$

$$P''(m,D) = \sum_{k=m}^{\infty} \frac{V(k, m+1)}{k!} \quad (4.4.5)$$

It can be noted from (4.4.3) that, for the computation of $V(n,m)$, only two elements of $V(n-1)$ (viz $V(n-1,m)$ and $V(n-1,m-1)$) are required. Hence, only two elements of $V(m)$ need to be stored at any stage of the computation. (We call it stages because starting with $P''(0,D)$, first $P''(1,D)$, then $P''(2,D)$ etc will be computed in that order). This implies that the storage requirements for $V(m)$ is minimal even though we have defined them as infinite dimensional matrix arrays. Further, because of the factor " $k!$ " in the denominator of (4.4.5), only finite terms need to be considered for the computation of $P''(m,D)$. The maximum value of m upto which $P''(m,D)$ is significant depends of course on the traffic arrival rate Λ'' and D . When the elements of the

matrix \underline{A}^D is less than 1, this procedure converges with reasonable speed

It can also be noted that if the diagonal elements of the arrival rate matrix of the MMPP had been distinct no further reduction in the computation time is possible. However, in order to keep track of the phases of MMPP 1 and MMPP 2 simultaneously, we defined MMPPs, MMPP 1 and MMPP 2 whose diagonal elements are not all distinct. In this case some simplifications are possible and are considered next.

The $(1, j)^{th}$ element of $P''(m, t)$ can be rewritten in terms of the parameters of MMPP 1 and MMPP 2 as follows

$$[P''(m, t)]_{1, j} = P[N''(t)=m, \underline{j}(t)=j | \underline{j}(0)=1] \quad (4.4.6)$$

$$= P[N''(t)=m, J'(t)=j', J''(t)=j'' | J'(0)=1', J''(0)=1''] \quad (4.4.7)$$

$$= P[N''(t)=m, J''(t)=j'' | J''(0)=1''] P[J'(t)=j' | J'(0)=1'] \quad (4.4.8)$$

$$= [\hat{P}''(m, t)]_{1'', j''} [e^{Q^{*'}t}]_{1', j'} \quad (4.4.9)$$

where $1 = (1'-1)M + 1''$ and $j = (j'-1)M + j''$. Here M denotes the total number of phases of MMPP 1. We obtain (4.4.8) from (4.4.7) using the fact that MMPP 1 and MMPP 2 are independent. $\hat{P}''(m, t)$ is the matrix $P(m, t)$ corresponding to MMPP2. The advantage of this observation is that, for computing $P''(m, t)$ it is enough if we find the matrices $\hat{P}''(m, t)$ whose dimension is small by a factor M , resulting in smaller computation time. For the computation of $\hat{P}''(m, t)$ the above recursive procedure can be used. The computation of $e^{Q^{*'}t}$ is straight forward and for 2×2 matrices, it is given by

$$e^{Q^{*'}t} = \begin{bmatrix} 1-r(t) & r(t) \\ \frac{r_2}{r_1} r(t) & 1 - \frac{r_2}{r_1} r(t) \end{bmatrix} \quad (4.4.10)$$

$$r(t) = \frac{r_1}{r_1+r_2} \left[1 - e^{-(r_1+r_2)t} \right]$$

where r_1, r_2 are the $(1, 2)^{th}, (2, 1)^{th}$ elements of Q^{*} .

It may also be noted that (4.4.8) can be written in matrix form as

$$P''(m,t) = e^{Q''t} \otimes \hat{P}''(m,t) \quad (4.4.11)$$

Similar simplifications could have also been done in the computation of busy period distribution of Q2. Busy period distribution of Q2 could have been expressed in terms of the parameters of MMPP 2 and MMPP 1. In order to keep the recursive procedure simple we have not done that here. However, in the case where the high priority queue size is finite and the traffic offered is high (the case studied in Chapter 7), we find that this approach helps in reducing the computation time and storage requirements significantly and is worth the additional complexity.

Finally we consider the computation of the matrices $P'(m,nD)$. One direct approach is to follow the same steps as that for the computation of $P''(m,D)$. However, even though the recursive procedure for $P'(m,nD)$ converges fast for $n=1$ when the elements of $\underline{\Lambda}'D$ are less than 1, it may not do so for higher values of n . The convergence problem is overcome as follows. Using the above recursive procedure $P'(m,D)$ is computed. To compute $P'(m,nD)$ for $n > 1$ another recursive procedure is used the details of which are considered next. We first note that the z transform of the matrices $P'(m,nD)$ corresponding to an MMPP can be expressed in terms of that of $P'(m,D)$. This can be verified by looking at the corresponding expressions given by

$$\mathcal{P}'(z,D) = \sum_{m=0}^{\infty} P'(m,D)z^m = e^{[\underline{\Lambda}'(z-1) + \underline{Q}^*]D} \quad (4.4.12)$$

$$\mathcal{P}'(z,nD) = \sum_{m=0}^{\infty} P'(m,nD)z^m = e^{[\underline{\Lambda}'(z-1) + \underline{Q}^*]nD} \quad (4.4.13)$$

$$= \left[\mathcal{P}'(z,D) \right]^n = \left[e^{[\underline{\Lambda}'(z-1) + \underline{Q}^*]D} \right]^n \quad (4.4.14)$$

Let the coefficient of (z^0, z^1, z^2, \dots) in $\{\mathcal{P}'(z,D)\}^n$ be represented as a matrix array $W(n)$ whose i^{th} element $W(n,i)$ gives the coefficient of z^{i-1} . From

(4.4.13), it can be noted that $W(n,1)$ is equal to $P'(1-1,nD)$. Let $W(0)$ be defined to be the matrix array with $W(0,m) = I \delta_{0m}$. The $(m+1)^{th}$ element of $W(n)$ denoted as $W(n,m+1)$ is equal to $P'(m,nD)$ and the elements of $W(n)$ can be obtained by convolving $W(n-1)$ with $W(1)$. Hence, $W(n,m+1)$ and $P'(m,nD)$ are given by

$$P'(m,nD) = W(n,m+1) = \sum_{k=1}^{m-1} W(1,k)W(n-1,m+1-k) \quad (4.4.15)$$

Hence knowing $W(1)$ or equivalently knowing the $P'(m,D)$ s, the $P'(m,nD)$ s can be recursively computed

4.5. AVERAGE DURATIONS OF THE BUSY PERIODS AND NUMBER OF CELLS SERVED DURING THE BUSY PERIODS OF Q1 AND Q2

We consider Q2 first. The computation of the average length of the busy period and the average number served during the busy period of Q2 can be carried out along the same lines as for the N/G/1 queue. It may be noted that the busy period of Q2 which starts with 1 customers is the sum of 1 first passage times of $Q''()$ i.e. the busy period of Q2 starts in level 1. It takes one first passage time of $Q''()$ for Q2 to reach level $i-1$. It takes another first passage time for it to reach level $i-2$ and so on until the level 0 is reached. Hence next, we consider the computation of these parameters in an interval of the first passage time.

It can be noted that the average number of customers served during the first passage time and its average duration can be found by computing the first moments of $\tilde{G}''(z,s)$ w.r.t z and s respectively. Let the j^{th} element of the vector μ'' denote the expected first passage time from $(i+1,j)$ to (i,j) for $i \geq 0$ in the semi-Markov process $Q''()$ and the j^{th} element of $\tilde{\mu}''$ denote the number of service completions during such a first passage from $(i+1,j)$ to

(1,j) Using Theorems 4 and 7 of Neuts [3] on the mean length of the first passage time and the mean number of customers served in the first passage time of a semi-Markov process, μ'' and $\tilde{\mu}''$ can be computed and is given by

$$\tilde{\mu}'' = \left. \frac{\delta \tilde{G}''(z, s)}{\delta z} \right|_{z=1, s=0} = [I - G'' + e g''] [I - A'' + (e - \beta'') g''] e \quad (4.5.1)$$

$$\mu'' = - \left. \frac{\delta \tilde{G}''(z, s)}{\delta s} \right|_{z=1, s=0} = [I - G'' + e g''] [I - A'' + (e - \beta'') g''] e \mu^{(1)''} \quad (4.5.2)$$

where g'' is the invariant probability vector of G'' and $A'' = \tilde{A}''(1, 0)$ Using (3.6.12) and (3.6.3), we get

$$A'' = \exp [Q^* D]$$

β'' is defined in section (4.2) and can be computed as-

$$\beta'' = \sum_{m=0}^{\infty} m \tilde{A}_m''(0) e = \left. \frac{\delta \tilde{A}''(z, s)}{\delta z} \right|_{z=1, s=0} \quad (4.5.3)$$

$\mu^{(1)''}$ denotes the average interdeparture time of cells from Q2 given that the previous departure left Q2 non-empty. It can be computed as-

$$\mu^{(1)''} e = \sum_{m=0}^{\infty} \int_0^{\infty} t dA_m''(t) e = - \left. \frac{\delta \tilde{A}''(z, s)}{\delta s} \right|_{z=1, s=0} \quad (4.5.4)$$

We next consider the evaluation of β'' . Using (3.6.5) in (4.5.3), we get

$$\beta'' = \sum_{m=0}^{\infty} m P''(m, D) e = \left. \frac{d \tilde{P}''(z, D)}{dz} \right|_{z=1} \quad (4.5.6)$$

Hence to find β'' , the first moment of $\mathcal{P}''(z, t)$ wrt z should be computed. This is a standard problem and the results are known for an MMPP (e.g. Heffes [9], Fischer [10]). However, since the intermediate steps used for computing the moment is useful for simplifying some of the later expressions we go through the steps involved in the evaluation as in Neuts [1]. We start with the Chapman-Kolmogorov equation satisfied by the counting functions $P''(m, t)$ of the MMPP $\underline{2}$ given by

$$\frac{d}{dt} P''(m, t) = P''(m, t)(\underline{Q}^* - \underline{\Lambda}'') + P''(m-1, t)\underline{\Lambda}'' \quad \text{for } m \geq 1, t \geq 0 \quad (4.5.6)$$

$$\frac{d}{dt} P''(0, t) = P''(0, t)(\underline{Q}^* - \underline{\Lambda}'') \quad t \geq 0$$

Let $V_1(t)$ and $V_0(t)$ be defined as follows -

$$V_1(t) = \sum_{m=0}^{\infty} m P''(m, t)$$

$$V_0(t) = \sum_{m=0}^{\infty} P''(m, t)$$

Multiplying both sides of (4.5.6) by m and adding the terms corresponding to $m=0, 1$, it can be verified that

$$\frac{d}{dt} V_1(t) = V_1(t)\underline{Q}^* + V_0(t)\underline{\Lambda}'' \quad (4.5.7)$$

Integrating (4.5.7) and multiplying it by e and noting the fact that $\underline{Q}^* e = 0$, we get

$$V_1(t)e = \int_0^t V_0(t) dt \underline{\Lambda}'' e = \int_0^t e^{\underline{Q}^* t} dt \underline{\Lambda}'' e \quad (4.5.8)$$

The integral on the RHS of (4.5.8) can be evaluated using Theorem (5.3.1) of Neuts[1] which states that for any irreducible infinitesimal generator matrix Q with stationary vector π , the matrix $e\pi - Q$ and $e\pi + Q$ are nonsingular and

$$\int_0^t e^{Qu} du = e\pi t + [I - e^{Qt}] [e\pi - Q]^{-1} \quad (4.5.9)$$

$$= e\pi t + [e^{Qt} - I] [e\pi + Q]^{-1} \quad (4.5.10)$$

In view of the nonsingularity of $e\pi + Q$ it also follows that

$$\pi = \pi(e\pi + Q)^{-1} \quad (4.5.11)$$

Let π be the invariant probability vector of \underline{Q}^* and let $V(t)$ be defined as

$$V(t) = e\pi t + [e^{\underline{Q}^* t} - I][e\pi + \underline{Q}^*]^{-1} \quad (4.5.12)$$

Using (4.5.10) and (4.5.12) in (4.5.9) we get

$$\beta'' = V_1(D)e = V(D)\underline{\Lambda}'' e \quad (4.5.13)$$

Finally, $\mu^{(1)''}$ can be computed, using (3.6.16) and (3.6.3) in (4.5.4), as

$$\mu^{(1)''} e = D e \quad (4.5.14)$$

Next we consider the computation of the average duration of the first busy period of Q2 and the number of customer served during this period. Towards this end, we define $\tilde{L}''(z,s)$ to be the joint transform of the number served and the duration of the first busy period of the server in Q2, given that at time 0 there are no customers in the system. The matrix of mass functions $L''(k,t)$ associated with $\tilde{L}''(z,s)$ is such that its (j,j') th entry is the conditional probability given $X''(0)=0$ and $J(0)=j$, that the first busy period is of duration less than or equal to t and consists of k services and that at the epoch where the BP of Q2 ends, the phase of MMPP $\underline{2}$ is j' . Considering all the possible ways in which the BP of Q2 can start, we get

$$\tilde{L}''(z,s) = \sum_{k=1}^{\infty} \tilde{U}_k''(0) \tilde{G}''^k(z,s) = \tilde{U}''[\tilde{G}''(z,s), 0] \quad (4.5.15)$$

Let $\tilde{\mu}_1''$ and μ_1'' be the $MN \times 1$ vectors whose i th elements denote respectively, the mean number served during and the mean duration of the first busy period of Q2 given that $X''(0)=0$ and $J''(0)=j$. It can be noted that $\tilde{\mu}_1''$ and μ_1'' can be computed by finding the first moments of $\tilde{L}''(z,s)$ w.r.t z and s respectively. The details of the computation of these moments and the simplifications that can be achieved are considered next. Differentiating $\tilde{L}''(z,s)$ w.r.t z and using the fact that G'' is stochastic we get

$$\begin{aligned} \tilde{\mu}_1'' = \sum_{k=1}^{\infty} \tilde{U}_k''(0) \left\{ \frac{\partial \tilde{G}''(z,s)}{\partial z} [G''(z,s)]^{k-1} e + G''(z,s) \frac{\partial \tilde{G}''(z,s)}{\partial z} [G''(z,s)]^{k-2} e \right. \\ \left. + [G''(z,s)]^{k-1} \frac{\partial \tilde{G}''(z,s)}{\partial z} e \right\} \Bigg|_{z=1, s=0} \end{aligned} \quad (4.5.16)$$

$$= \sum_{k=1}^{\infty} \tilde{U}_k''(0) \frac{\partial G''(z,s)}{\partial z} e + G'' \frac{\partial G''(z,s)}{\partial z} e + G''^2 \frac{\partial G''(z,s)}{\partial z} e + G''^{k-1} \frac{\partial G''(z,s)}{\partial z} e$$

$$= \sum_{k=1}^{\infty} \tilde{U}_k''(0) \sum_{v=1}^{k-1} G''^v [I - G'' + e g''] [I - G'' + e g'']^{-1} \tilde{\mu}'' \quad (4.5.17)$$

$$= \sum_{k=1}^{\infty} \tilde{U}_k''(0) [I - G''^k + k e g''] [I - G'' + e g'']^{-1} \tilde{\mu}'' \quad (4.5.18)$$

$$= [\tilde{U}''(1,0) - \tilde{U}''(G'',0) + \underline{\tilde{U}}''(1,0) e g''] [I - G'' - e g'']^{-1} \tilde{\mu}'' \quad (4.5.19)$$

$$\text{where } \underline{\tilde{U}}''(1,0) = \sum_{k=1}^{\infty} k \tilde{U}_k''(0) = \left. \frac{\delta \tilde{U}''(z,s)}{\delta z} \right|_{z=1, s=0} \quad (4.5.20)$$

It may be recalled that the matrix $\tilde{U}''(z,s)$ is defined in section (3.6). Similarly, differentiating (4.5.5) w.r.t. s and evaluating the resulting expression at $z=1$ and $s=0$ we get

$$\mu_1'' = [\tilde{U}''(1,0) - \tilde{U}''(G'',0) + \underline{\tilde{U}}''(1,0) e g''] [I - G'' - e g'']^{-1} \mu'' \quad (4.5.21)$$

It may be observed that the corresponding expressions obtained for the N/G/1 queue in Ramaswami [2] differs from (4.5.19) and (4.5.21) in the third term within the first square bracket. This is because in the N/G/1 queue the time when the first batch of arrival occurs at the queue is identical to the time when the busy period starts, however in Q2 of our system, the BP does not start the moment a cell arrives at an empty Q2 if Q1 is non-empty at that instant.

We consider the evaluation of $\tilde{\mu}_1''$ and μ_1'' using the results of Sec. 3.6. Using (3.6.18) and (3.6.3), we get

$$\tilde{U}''(1,0) = [-R''(0)]^{-1} \left\{ p_0' \underline{\Lambda}'' + (1-p_0') \frac{1}{D} \int_{w=0}^D \underline{\Lambda}'' dw e^{-\underline{Q}''^* w} \right\} \quad (4.5.22)$$

$$\tilde{U}''(G'',0) = [-R''(0)]^{-1} \left\{ p_0' \underline{\Lambda}'' + (1-p_0') \frac{1}{D} \int_{w=0}^D \underline{\Lambda}'' dw e^{R''(G'')w} \right\} \quad (4.5.23)$$

Let π be the invariant probability vector of \underline{Q}''^* . Then, using (4.5.10) and

(4 5 12) in (4 5 22), we get

$$\tilde{U}''(1,0) = [\underline{\Lambda}'' - \underline{Q}^*]^{-1} \underline{\Lambda}'' \left[p_0' I + (1-p_0') \frac{1}{D} \nu(D) \right] \quad (4 5 24)$$

For evaluating the integral in (4 5 23), it may be noted that the matrix $R''(G'')$ satisfies the properties of an irreducible, infinitesimal generator matrix. To illustrate this point, let us compute $R(G)$ corresponding to a 2-state MMPP/D/1 queue. Let (Λ, Q^*) be the arrival rate and infinitesimal generator matrices of the 2-phase MMPP. Denoting the $(i,j)^{th}$ elements of G as G_{ij} and the i^{th} diagonal element of Λ as λ_i we get

$$\Lambda G - \Lambda = \begin{bmatrix} \lambda_1 G_{11} & \lambda_1 G_{12} \\ \lambda_2 G_{21} & \lambda_2 G_{22} \end{bmatrix} - \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} -\lambda_1 G_{12} & \lambda_1 G_{12} \\ \lambda_2 G_{21} & -\lambda_2 G_{21} \end{bmatrix}$$

Hence $\Lambda G - \Lambda$ satisfies the properties of an infinitesimal generator matrix (viz row sums zero, diagonal entries negative and nondiagonal elements non-negative). As Q^* is an infinitesimal generator matrix, $R(G) = \Lambda G - \Lambda + Q^*$ also forms an infinitesimal generator matrix. Hence (4 5 10) can be used to compute the integral in (4 5 22). Let $\tilde{\pi}''$ denote the invariant probability vector of $R''(G'')$. Using (4 5 10) in (4 5 22)

$$\begin{aligned} \tilde{U}''(G'',0) &= [\underline{\Lambda}'' - \underline{Q}^*]^{-1} \underline{\Lambda}'' \left\{ p_0' I + (1-p_0') \left[e^{\tilde{\pi}''} + \frac{1}{D} [e^{R''(G'')D} - I] \right. \right. \\ &\quad \left. \left. [e^{\tilde{\pi}''} + R''(G'')]^{-1} \right] \right\} G'' \end{aligned} \quad (4.5 25)$$

Using (3 6 19), (4 5 10) and (4 5 12) in (4 5 20) and noting that $\tilde{U}''(1,0)e = e$, we get

$$\begin{aligned} \tilde{U}''(1,0)e &= \tilde{U}''(1,0)e + [-R''(0)]^{-1} (1-p_0') \frac{1}{D} \int_{w=0}^D dw \underline{\Lambda}'' \frac{\delta \mathcal{P}''(z,w)}{\delta z} e \Big|_{z=1} \\ &= e + [-R''(0)]^{-1} (1-p_0') \frac{1}{D} \int_{w=0}^D dw \underline{\Lambda}'' \left\{ e^{\pi'' w} + [e^{Q^* w} - I] [e^{\pi''} + Q^*]^{-1} \right\} \underline{\Lambda}'' e \end{aligned}$$

$$= \mathbf{e} + [-\mathbf{R}''(0)]^{-1}(1-p_0')\underline{\Lambda}''\left\{\mathbf{e}\pi''\frac{D}{2} + \left[\frac{1}{D}\mathbf{V}(D)-\mathbf{I}\right][\mathbf{e}\pi'' + \mathbf{Q}^*]^{-1}\right\}\underline{\Lambda}''\mathbf{e} \quad (4.5.26)$$

We now consider the other queue Q1. The computation of the average length of the busy period and the average number served during the busy period of Q1 is carried out along the same lines as for Q2. Let \mathbf{g}' denote the invariant probability vector of \mathbf{G}' . Let the j^{th} element of the vector μ' denote the expected first passage time from $(i+1, j)$ to (i, j) for $i \geq 0$ in the semi-Markov process $\mathbf{Q}'(\cdot)$ and the j^{th} element of $\tilde{\mu}'$ denote the number of service completions during such a first passage from $(i+1, j)$ to (i, j) . Further let $\tilde{\mu}'_1$ and μ'_1 be the $MN \times 1$ vectors whose i^{th} elements denote respectively, the mean number of cells served during the first busy period of Q1 and the mean duration of the first busy period of Q1 given that $\mathbf{X}'(0)=0$ and $\mathbf{j}'(0)=\mathbf{j}$.

The expressions for $\tilde{\mu}'$, μ' , $\tilde{\mu}'_1$ and μ'_1 can be obtained by replacing the parameters of Q2 by those of Q1 in (4.5.1)-(4.5.4) and (4.5.15)-(4.5.21). To evaluate these expressions, we obtain the expressions for $\mu^{(1)'}\mathbf{e}$, β' , $\tilde{\mathbf{U}}'(1,0)$, $\tilde{\mathbf{U}}'(\mathbf{G}',0)$ and $\tilde{\mathbf{U}}'(1,0)\mathbf{e}$ using the results of Sec 3.6. Using this, $\mu^{(1)'}\mathbf{e}$ can be found by replacing the parameters of Q2 by those of Q1 in (4.5.4). Using (3.6.17) and noting that $\exp\{\mathbf{R}'(1)kD\}$ is stochastic, we get

$$\begin{aligned} \mu^{(1)'}\mathbf{e} &= -\left.\frac{\delta \tilde{\mathbf{A}}'(z,s)}{\delta s}\mathbf{e}\right|_{z=1,s=0} \\ &= \sum_{k=1}^{\infty} C(k-1)kDe^{[\mathbf{R}'(1)]kD}\mathbf{e} = \sum_{k=1}^{\infty} C(k-1)kD\mathbf{e} \end{aligned} \quad (4.5.27)$$

β' is computed using (3.6.15) and (4.5.10) and (4.5.12) and is given by

$$\beta' = \sum_{m=0}^{\infty} m \tilde{\mathbf{A}}'_m(0) \mathbf{e} = \left.\frac{\delta \tilde{\mathbf{A}}'(z,s)}{\delta z}\mathbf{e}\right|_{z=1,s=0} \quad (4.5.28)$$

$$= \sum_{k=1}^{\infty} C(k-1)V(kD)\underline{\Lambda}'\mathbf{e} \quad (4.5.29)$$

To simplify the notation, we define a matrix function $\mathfrak{F}(\mathbf{G}')$ as follows-

$$\mathcal{F}(G') = \sum_{\iota=0}^{\infty} F(\iota) e^{[R'(G')]\iota D} \quad (4.5.30)$$

The matrix obtained by replacing G' by I in $\mathcal{F}(G')$ is denoted as $\mathcal{F}(I)$

Using (4.5.10), (4.5.12) and (4.5.30) in (3.6.22), it can be verified that

$$\tilde{U}'(1,0) = [\Lambda' - Q^*]^{-1} \Lambda' \left[p_0'' I + (1-p_0'') \frac{1}{D} V(D) \mathcal{F}(I) \right] \quad (4.5.31)$$

$$\tilde{U}'(G',0) = [\Lambda' - Q^*]^{-1} \Lambda' \left\{ (1-p_0'') \frac{1}{D} \sum_{n=0}^{\infty} \int_{w=0}^D P'(n,w) dw \mathcal{F}(G') G'^{n+1} + G' p_0'' \right\} \quad (4.5.32)$$

$$\begin{aligned} \tilde{U}'(1,0) &= \sum_{k=1}^{\infty} k \tilde{U}'_k(0) = \left. \frac{\delta \tilde{U}'(z,s)}{\delta z} \right|_{z=1, s=0} \\ &= e + [-R'(0)]^{-1} (1-p_0'') \Lambda' \left\{ \frac{1}{D} V(D) \sum_{\iota=0}^{\infty} F(\iota) V(\iota D) \Lambda' e \right. \\ &\quad \left. + e \pi \frac{D}{2} + \left[\frac{1}{D} V(D) - I \right] [e \pi + Q^*]^{-1} \right\} \Lambda' e \end{aligned} \quad (4.5.33)$$

Finally, we consider some of the approximations that can be used in the above expressions. It may be noted that the elements of the matrices $C(k)$ and $F(\iota)$ become negligible for large values of k and ι respectively and hence only finite terms need to be considered for the computation β' and $\mathcal{F}(G')$. The maximum values of k and ι depends on the traffic offered at Q2 and is expected to be small. Similarly, as $P'(n,w)$ becomes small for large values of n , again only finite terms need to be considered in (4.5.32). The integral in (4.5.32) has to be numerically computed, it may be precomputed and stored for being repeatedly used in the successive iteration of the QLDs.

4.6. AVERAGE DURATION OF THE BUSY CYCLES OF Q1 AND Q2

We define the busy cycle of Q2 to be the time between the successive visits of $Q''()$ to the level 0. In other words it is the sum of the vacation interval and the busy period of the server as viewed from Q2. It may be

recalled that the server goes on vacation from Q2 as and when it becomes empty. The server resumes service for Q2 when it finds Q2 to be non-empty again and the ongoing service, if any, for the Q1 cell is completed. The vacation period consists of the idle period (when no cell is found in Q2) and the residual service time of a Q1 cell. For the computation of the average duration of the busy cycle, we define matrices $K_0''(n,t)$ whose (i,j) th element gives the joint probability that the busy cycle is of length $\leq t$, consists of n customer services and ends with the MMPP $\underline{2}$ phase as j given that at time 0 the phase is i . We denote the double transform of $K_0''(n,t)$ as $\tilde{K}_0''(z,s)$. An expression for $\tilde{K}_0''(z,s)$ can be obtained as follows. Busy cycle is the sum of 2 random variables viz vacation period and busy period. The number of cells served again is the sum of two random variables viz the number that arrived in the vacation interval and the number that arrived during the busy period. Hence $\tilde{K}_0''(z,s)$ is given by

$$\tilde{K}_0''(z,s) = \sum_{m=0}^{\infty} \tilde{U}_m''(s) \tilde{G}''^m(z,s) \quad (4.6.1)$$

Let the i th element of the vector $\hat{\mu}''$ denote the average length of the busy cycle given that the busy cycle started in phase i . Using (4.6.1) $\hat{\mu}''$ can be obtained as follows

$$\begin{aligned} \hat{\mu}'' &= - \left. \frac{\partial \tilde{K}_0''(z,s)}{\partial s} e \right|_{z=1, s=0} \\ &= - \left\{ \sum_{m=0}^{\infty} U_m''(0) \frac{d}{ds} [\tilde{G}''(1,s)] e \right|_{s=0} + \sum_{m=0}^{\infty} \frac{d}{ds} [\tilde{U}_m''(s)] e \Big|_{s=0} \right\} \\ &= \mu_1'' + \mu_2'' \end{aligned} \quad (4.6.2)$$

where the i th elements of the vectors μ_1'' and μ_2'' denote the average length of the vacation interval and the busy period starting in phase i .

Computation of the vector μ_1'' has been considered in detail in section (4.5). Hence, we obtain an expression for μ_2'' in this section. Towards this

end, we need an expression for the derivative of the inverse of a matrix function $Y(t)$. Using the following matrix identity and differentiating it, we get

$$[Y(t)][Y(t)]^{-1} = I$$

$$\frac{d}{dt} \left\{ [Y(t)][Y(t)]^{-1} \right\} = 0$$

$$\left\{ \frac{d}{dt}[Y(t)] \right\} [Y(t)]^{-1} + [Y(t)] \frac{d}{dt}[Y(t)]^{-1} = 0$$

Thus

$$\frac{d}{dt}[Y(t)]^{-1} = - [Y(t)]^{-1} \left\{ \frac{d}{dt}[Y(t)] \right\} [Y(t)]^{-1} \quad (4.6.3)$$

Using (3.6.18), (3.6.3)-(3.6.4) and (4.6.3) in the second term of (4.6.2) we get

$$\begin{aligned} \mu_2'' &= - \sum_{m=0}^{\infty} \frac{d}{ds} [\tilde{U}_m''(s)] e \Big|_{s=0} = - \frac{d}{ds} [\tilde{U}''(1, s)] e \Big|_{s=0} \\ &= [\Lambda'' - Q^*]^{-1} e + (1-p_0') \frac{D}{2} e \end{aligned} \quad (4.6.4)$$

It can be noted that the first term of (4.6.4) is equal to the average idle period of Q2 when Q2 is empty. The second term gives the contribution from the residual service time of Q1. If the service discipline had been FCFS the second term would have been absent.

Busy cycle of Q1 is defined along the same lines. In this case vacation interval consists of 3 parts viz the initial idle period, the RST of the Q2 cell undergoing service and the ABP of Q2 that follows. We define for Q1 the vectors $\hat{\mu}'$, μ_1' and μ_2' whose i th elements denote the average length of the busy cycle, the vacation interval and the busy period starting in phase i respectively. These vectors can be obtained using (4.6.1)-(4.6.3) by replacing the parameters of Q2 by those of Q1.

As the expression for μ_1' has been obtained in section (4.5), We next

consider the evaluation of μ_2' Using (3.6.24), (3.6.3)-(3.6.4) and (4.6.3) in the second term of (4.6.2) we get

$$\begin{aligned}\mu_2' &= - \sum_{m=0}^{\infty} \frac{d}{ds} [\tilde{U}_m'(s)] e \Big|_{s=0} = - \frac{d}{ds} [\tilde{U}'(1,s)] e \Big|_{s=0} \\ &= [\Lambda' - Q^*]^{-1} e + (1-p_0'') \frac{D}{2} e + (1-p_0'') [\Lambda' - Q^*]^{-1} \Lambda' \frac{1}{D} \mathcal{V}(D) \sum_{\iota=0}^{\infty} F(\iota) \iota D e \quad (4.6.5)\end{aligned}$$

It can be observed that the second and third terms of (4.6.5) arise due to the residual service time(RST) of a Q2 cell and the additional busy period that might follow this RST respectively

Finally, we consider the computation of the stationary vector of $\tilde{K}_0''(1,0)$ and $\tilde{K}_0'(1,0)$ denoted as k_0'' and k_0' respectively. It may be noted that $\tilde{K}_0''(1,0) = U''(G'',0)$ and $\tilde{K}_0'(1,0) = U'(G',0)$. Hence k_0'' and k_0' can be computed using (4.5.25) and (4.5.32). However, for the N/G/1 queue (which is referred to as BMAP/G/1 in Lucantoni[11]), simple formulae for the computation of the vectors x_0 , y_0 and k were obtained in terms of the stationary vector g of the matrix G . This results in significant reduction in storage and computational requirements as the expressions analogous to those of section (4.5) and (4.6) which were required to be evaluated using the approach of Ramaswami[2] could altogether be dispensed with. We have so far been following closely the approach of [2]. We now examine whether the simplifications suggested in [11] carry over to the present problem.

Since the inter departure time of cells from Q2 (IDT 2), given that at the previous departure left Q2 non-empty, does not depend on the phase of the arrival process, the IDT 2 is a scalar stochastic process. Let the c d f of IDT 2 be denoted as $\mathcal{H}''(t)$. Using (3.4.1) we get

$$\mathcal{H}''(t) = u(t-D) \quad (4.6.6)$$

Now, proceeding along the same lines as in [11], it can be shown that the

matrix $G''(z,s)$ satisfies the equation given by

$$\tilde{G}''(z,s) = z \int_0^{\infty} e^{-st} e^{\{R''(\tilde{G}''(z,s))\}t} d\mathcal{H}''(t) \quad (4.6.7)$$

Substituting $s=0$ and $z=1$ we get

$$G'' = \int_0^{\infty} e^{\{R''(G'')\}t} d\mathcal{H}''(t) \quad (4.6.8)$$

In section (4.5) we have shown that $R''(G'')$ is the infinitesimal generator of an irreducible Markov process. Let w be its stationary vector i.e. $w e = 1$ and $w R''(G'') = 0$. Using (4.6.8) it can be verified that w is also the stationary vector of G'' . Since the stationary vector of G'' is unique, w has to be equal to g'' . Hence we get

$$g'' R''(G'') = g'' (\underline{\Lambda}'' G'' - \underline{\Lambda}'' + \underline{Q}''^*) = 0 \quad (4.6.9)$$

Next, we consider the computation of k_0'' . For an MMPP/D/1 queue with FCFS discipline, the application of the results of [11], yields k_0'' to be equal to $g''(\underline{\Lambda}'' - \underline{Q}''^*)$. Analogous to this, we examine whether $g''(\underline{\Lambda}'' - \underline{Q}''^*)$ is equal to k_0'' . Using (4.5.25) and (4.6.9) we get

$$\begin{aligned} g''(\underline{\Lambda}'' - \underline{Q}''^*) K''(1,0) &= g''(\underline{\Lambda}'' - \underline{Q}''^*) [\underline{\Lambda}'' - \underline{Q}''^*]^{-1} \underline{\Lambda}'' \left[p_0' G'' \right. \\ &\quad \left. + (1-p_0') \left\{ eg'' + \frac{1}{D} [e^{R''(G'')D} - I][eg'' + R''(G'')]^{-1} G'' \right\} \right] \\ &= g'' \underline{\Lambda}'' p_0' G'' + g'' \underline{\Lambda}'' (1-p_0') \left\{ eg'' + \frac{1}{D} [e^{R''(G'')D} - I][eg'' + R''(G'')]^{-1} G'' \right\} \\ &= g''(\underline{\Lambda}'' - \underline{Q}''^*) p_0' + g'' \underline{\Lambda}'' (1-p_0') \left\{ eg'' + \frac{1}{D} [e^{R''(G'')D} - I][eg'' + R''(G'')]^{-1} G'' \right\} \quad (4.6.10) \end{aligned}$$

It can be noted from (4.6.10) that for the FCFS discipline (single priority case) $p_0' = 1$ and $g''(\underline{\Lambda}'' - \underline{Q}''^*) = k_0''$. Otherwise $g''(\underline{\Lambda}'' - \underline{Q}''^*) \neq k_0''$. This is a major break down and is a departure from the BMAP/G/1 queue behaviour. Hence it can be concluded that simplifications of [11] are not applicable for Q2. Since the inter departure time of cells from Q1 (IDT 1), given that at the previous

departure left Q1 non-empty, depends on the phase of the arrival process, the IDT 1 is a vector stochastic process and $\tilde{G}'(z,s)$ does not satisfy the equation analogous to (4.6.7). Hence no simplification is possible also for Q1

REFERENCES

- 1 M F Neuts, "Structured stochastic matrices of the M/G/1 type and their applications, Marcel Dekker, New York, 1989
- 2 V Ramaswami, "The N/G/1 queue and its detailed analysis", Adv Appl Prob 12(1980), pp 222-261
- 3 M F Neuts, "Moment formulas for the Markov Renewal branching process", Adv Appl Prob 8, pp 690-711, 1976
- 4 L Takacs, "Introduction to the theory of queues", Oxford University Press, New York, 1962
- 5 W Feller, "Introduction to probability theory and its applications", Second Edition, John Wiley & sons, Inc New York, 1966
- 6 D M Lucantoni and V Ramaswami, "Efficient algorithms for solving the non-linear equations arising in phase type queues", Commn Statist Stochastic Models 1, 1985, pp 29-51
- 7 V Ramaswami, "The busy period of queues which have a matrix-geometric steady state probability vector", Opsearch 19, 1982 pp 238-261
- 8 M F Neuts, "Matrix - Geometric Solutions in Stochastic Models An Algorithmic Approach", The Johns Hopkins University Press, Baltimore, 1981
- 9 H Heffes and D M Lucantoni, " A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance", IEEE J Selected Areas in Commn, 4(6), 1986, pp 856-868
- 10 W Fischer, K Meier-Hellstern, "The Markov modulated Poisson process (MMPP) cookbook", Performance evaluation 18, 1993, pp 149-171

- 11 D M Lucantoni, "New results on the single server queue with a batch Markovian arrival process", *Commn Statist Stochastic Models* 7, 1991, pp 1-46

CHAPTER 5

EVALUATION OF THE QUEUE LENGTH DENSITIES OF Q1 AND Q2

5.1. INTRODUCTION

In this chapter, we present the details of the evaluation of the QLDs of Q1 and Q2 using numerical computations. Validations of these results using simulations are also discussed. We have introduced the exact model for the numerical computation in Sec. 3.2. The evaluation of some of the parameters required for computing the QLDs using this model have been discussed in Chapters 3 and 4. In this chapter, we consider first the computation of the remaining set of parameters required for finding the QLDs. Some numerical approaches for computing the QLDs are then discussed. This is followed by a discussion on an approximate model which is computationally and storage wise efficient for computing the QLDs. Some details on two simulation models for the computation of the busy period distribution of the higher priority queue and another model for computing the QLDs are presented next. Some details on the numerical computation of the QLDs for some typical examples and the approximations made for this purpose are discussed. Finally, the results obtained through the numerical computation are presented for a number of examples and compared with those obtained using simulations.

5.2. EXPRESSIONS FOR x'_0 AND x''_0

It may be recalled that the i^{th} element of the vector x''_0 denotes the joint probability that a departure from Q2 leaves Q2 empty and the process MMPP 2 in phase 1. Similarly, the i^{th} element of the vector x'_0 denotes the joint probability that a departure from Q1 leaves Q1 empty and MMPP 1 in phase 1. Computation of x'_0 and x''_0 is a crucial step in the computation of the QLDs

of Q1 and Q2. For example, knowing x_0'', x_1'' for $i=1,2,3$ can be recursively computed as discussed in Sec 5.6. Moreover, for the computation of p_0' and p_0'' , x_0' and x_0'' must be known. We consider Q2 first, for this, x_0'' can be evaluated by considering the Markov Renewal Process(MRP) $Q''()$ at its successive visits to the set $(0,j)$. We shall follow the approach given in Neuts [1].

It is obvious that whenever $Q''()$ visits the state $(0,j)$, the SMC $Q''(\omega)$ also visits the state $(0,j)$. Hence we shall concentrate on the successive visits of $Q''()$ to the set 0. We shall say that $\xi_n'' = j$, $1 \leq j \leq MN$, if the state entered at the n^{th} visit to the state 0 is $(0,j)$. Let the random variable φ_n'' , $n \geq 2$ denote the number of transitions between the $(n-1)^{\text{th}}$ and n^{th} visits to the set 0. As the embedded points of $Q''()$ are chosen to be the departure instants of Q2 cells, φ_n'' also gives the number of customers served between the adjacent visits. Let θ_n'' denote the length of time between $(n-1)^{\text{th}}$ and the n^{th} visits to the level 0. We assume φ_0'' and θ_0'' to be 0. Initially, MRP $Q''()$ will not be in the set 0, hence, let φ_1'' and θ_1'' be defined as the number of transitions and the time required to reach the set 0 for the first time.

With this notation, it can be observed that for every $n \geq 2$, the pairs $(\varphi_1'', \theta_1'')$, $(\varphi_2'', \theta_2'')$, $(\varphi_3'', \theta_3'')$, ..., $(\varphi_n'', \theta_n'')$ are conditionally independent given the random variables ξ_0'' , ξ_1'' , ξ_2'' , ..., ξ_{n-1}'' . In other words, the values of $(\varphi_{n+1}'', \theta_{n+1}'')$ depends only on ξ_n'' and not on the values of $(\varphi_n'', \theta_n'')$. Hence $\{(\xi_n'', \varphi_n''), n \geq 0\}$ and $\{(\xi_n'', \theta_n''), n \geq 0\}$ form semi-Markov sequences respectively with the state spaces $\{1,2, \dots, MN\} \times \{0,1, \dots\}$ and $\{1,2, \dots, MN\} \times [0,\infty)$.

When $Q''()$ is recurrent, both Markov renewal sequences possess proper transition probability matrices. It may now be observed that the transition probabilities $P\{\xi_n'' = j', \varphi_n'' = k, \theta_n'' \leq t | \xi_{n-1}'' = j\}$ for $1 \leq j, j' \leq MN$, $k \geq 1$, $t \geq 0$ are essentially the same as the $(j,j')^{\text{th}}$ element of the matrices $K_0''(n,t)$ defined

in Sec 4.5 Hence the mean recurrence time (MRT) of the state $(0,j)$ in the Markov chain $Q''(\omega)$ is the same as that of the MRT of $(0,j)$ in the Markov renewal process of the lattice type $\tilde{K}_0''(z,0) = \tilde{L}''(z,0)$

By applying Theorem 2.11, pp 196, of Hunter [2] on the mean first passage time of a Markov Renewal process, the MRT of the state $(0,j)$, denoted as $m(0,j)$ is given by

$$m(0,j) = (k_0'' \tilde{\mu}_1'') [(k_0'')_j]^{-1} \quad (5.2.1)$$

where $(k_0'')_j$ is the j^{th} element of the invariant probability vector of $\tilde{K}_0''(1,0)$ and $x''(0,j)$ is the inverse MRT of the state $(0,j)$ in $Q''(\omega)$ $\tilde{\mu}_1''$ is given by (4.5.19) Hence x_0'' is given by

$$x_0'' = (k_0'' \tilde{\mu}_1'')^{-1} k_0'' \quad (5.2.2)$$

Let k_0' be the invariant probability vector of $\tilde{K}_0'(1,0)$ It can be shown that

$$x_0' = (k_0' \tilde{\mu}_1')^{-1} k_0' \quad (5.2.3)$$

$\tilde{\mu}_1'$ is obtained by using (4.5.31) - (4.5.33) in (4.5.19)

5.3. COMPUTATION OF x_1'' AND x_1'

Iterative schemes like the Block Gauss-Seidel scheme (see for e.g. Neuts [1]) for the computation of the QLDs of Q1 and Q2 require x_0'' , x_1'' , x_0' and x_1' for starting the iteration. Hence we consider the computation of x_1'' and x_1' . It may be noted that the steps used here may be used to compute x_1' and x_1'' for any value of $i \geq 1$. However the computational complexity becomes very high for large values of i .

It may be noted that the vector x_1'' denotes the joint probability that a departure from Q2 leaves behind one cell in Q2 and the MMPP $\underline{2}$ in phase i . Analogous to that in Sec 4.5 we define the matrices $K_1''(n,t)$ as follows. $K_1''(n,t)$ is an $MN \times MN$ matrix whose $(i,j)^{\text{th}}$ element denotes the probability that starting in $(1,1)$, the Markov renewal process $Q''(\cdot)$ returns for the first time

to the set 1 in exactly n steps at or before time t and that the phase of MMPP $\underline{2}$ at the epoch of such a first return is j . An expression for the double transform of $K_1''(n,t)$, denoted as $\tilde{K}_1''(z,s)$ can be written down as in Ramaswami [3] by considering the various possible ways in which the first return to level to 1 can occur starting from level 1. Firstly, after serving the current cell, Q2 may become empty. Following this there may be r busy periods of duration D sec each followed by r idle periods. During these periods, level 1 is not visited. Finally there may a BP which lasts longer than D sec in which case the level 1 will be definitely visited. Hence the first return time in this case is the sum of 1 cell service time, r busy cycles of Q2 in each of which only one cell is served and the time to taken to visit level 1 starting from level 0. In the second case, starting from level 1, $Q''()$ may return to level 1 without visiting level 0. This implies that during the service time that follows level 1, $v(\geq 1)$ customers arrive and hence before returning to level 1, $(v-1)$ first busy periods (i.e the time taken for $Q''()$ to go from level $i+1$ to i) will precede. Considering these two chain of events we get

$$\begin{aligned}
 \tilde{K}''(z,s) &= z\tilde{A}_0''(s) \sum_{r=0}^{\infty} z[\tilde{B}_0''(s)]^r \sum_{v=1}^{\infty} z\tilde{B}_v''(s) [\tilde{G}''(z,s)]^{v-1} + \sum_{v=1}^{\infty} z\tilde{A}_v''(s) [\tilde{G}''(z,s)]^{v-1} \\
 &= z^2\tilde{A}_0''(s) [I - z\tilde{B}_0''(s)]^{-1} \sum_{v=1}^{\infty} \tilde{B}_v''(s) [\tilde{G}''(z,s)]^{v-1} \\
 &\quad + \sum_{v=1}^{\infty} z\tilde{A}_v''(s) [\tilde{G}''(z,s)]^{v-1}
 \end{aligned} \tag{5.3.1}$$

Let k_1'' denote the stationary vector of $\tilde{K}_1''(1,0)$, then proceeding along the same lines as for the computation of x_0'' and using Theorem 3.2.11 of Hunter[2], we get

$$x_1'' = (k_1'' \tilde{k}_1'')^{-1} k_0'' \tag{5.3.2}$$

where $\tilde{k}_1'' = (\delta/\delta z)\tilde{K}''(z,0)$ is the mean number of steps in a first passage from 1 to itself and can be computed using (5.3.1). Simplifying the resulting expression along the same lines as for the computation of μ'' in Sec. 4.5 we get

$$\begin{aligned} \tilde{k}_1'' &= e + A_0'' [I - B_0'']^{-1} e + \left\{ A_0'' [I - B_0'']^{-1} \left\{ \sum_{v=1}^{\infty} B_v'' - \sum_{v=1}^{\infty} B_v'' G''^{v-1} + \sum_{v=2}^{\infty} (v-1) B_v'' e g'' \right\} \right. \\ &\quad \left. + (A'' - A_0'') - \sum_{v=1}^{\infty} A_v'' G''^{v-1} + \sum_{v=2}^{\infty} (v-1) A_v'' e g'' \right\} (I - G'' + e g'')^{-1} \mu'' \end{aligned} \quad (5.3.3)$$

where $G'' = \tilde{G}''(1,0)$ and g'' is the invariant probability vector of G'' . Similarly, the i^{th} element of the vector x_1' denotes the joint probability that a departure from Q1 leaves behind one cell in Q1 and the MMPP $\underline{1}$ in phase i . It can be verified that x_1' can be computed using (5.3.1) and (5.3.3) by replacing the parameters of Q2 by those of Q1.

5.4. MOMENTS OF THE QUEUE LENGTHS OF Q1 AND Q2

The evaluation of the QLDs of the queues of the "M/G/1 type", using the numerical methods, requires the infinite dimensional matrices like $Q''()$ and $Q'()$ to be truncated appropriately. The truncation may be carried out based on the value of the mean and variance of the queue lengths (see for e.g. [3]) and hence the computation of the QLDs may have to be preceded by the computation of the first and second moments of the queue lengths. Further, when the traffic offered to Q1 becomes close to the capacity of the server, the computational complexity and storage required for the computation of the QLD of Q1 may become prohibitively high and one may have to get limited to the knowledge of the first few moments of the queue lengths. Finally, computation of the moments of the queue length will enable the computation of the average queueing delays of cells arriving at Q1. Hence we consider the computation of the moments of the queue lengths in some detail. It may be noted that for the

queues of the "M/G/1 type", the steps involved in the computation of these moments is now standard and is given in Neuts[1]. We go through these steps for the present problem in some detail and indicate the simplifications that can be carried out.

First we obtain an expression for $X''(z)$, the z transform of x''_n . Let B''_m , A''_m and P''_m denote the matrices $\tilde{B}''_m(\omega)$, $\tilde{A}''_m(\omega)$ and $P''(m,D)$ respectively. Then using the system of equations

$$Q''(\omega) x'' = x''$$

$$x'' e = 1$$

$$x'' = [x''_0, x''_1, x''_2, \dots]$$

$$e = [e^T, e^T, e^T, \dots]^T$$

we get

$$x''_1 = x''_0 B''_1 + \sum_{k=1}^{1+1} x''_k A''_{1-k+1} \quad (5.4.1)$$

$$X''(z) = \sum_{i=0}^{\infty} x''_i z^i = x''_0 \sum_{i=0}^{\infty} z^i B''_1 + \sum_{i=0}^{\infty} z^i \sum_{k=1}^{1+1} x''_k A''_{1-k+1} \quad (5.4.2)$$

The first term of (5.4.2) can be simplified by using (3.3.6) and collecting the like terms as follows

$$\begin{aligned} B''(z) &= \sum_{i=0}^{\infty} z^i B''_1 = \sum_{i=0}^{\infty} z^i \sum_{k=1}^{1+1} U''_k P''_{1-k+1} \\ &= \frac{1}{z} \left\{ U''_1 z P''_0 + U''_1 z (P''_1 z) + U''_2 z^2 P''_0 \right. \\ &\quad + U''_1 z (P''_2 z^2) + U''_2 z^2 (P''_1 z) + U''_3 z^3 P''_0 + \\ &\quad \left. + U''_1 z (P''_3 z^3) + U''_2 z^2 (P''_2 z^2) + U''_3 z^3 P''_1 z + U''_4 z^4 P''_0 + \dots \right\} \\ &= \frac{1}{z} \left\{ U''_1 z \left[P''_0 + P''_1 z + P''_2 z^2 + P''_3 z^3 + P''_4 z^4 + \dots \right] \right. \\ &\quad \left. + U''_2 z^2 \left[P''_0 + P''_1 z + P''_2 z^2 + P''_3 z^3 + P''_4 z^4 + \dots \right] + \dots \right\} \end{aligned}$$

$$= \frac{1}{z} \left\{ U''(z,0) \mathcal{P}''(z,D) \right\} \quad (5.4.3)$$

Similarly it can be shown that

$$\sum_{i=0}^{\infty} z^i \sum_{k=1}^{i+1} x_k'' A_{i-k+1}'' = \frac{1}{z} \left[X''(z) - x_0'' \right] \tilde{A}''(z,0) \quad (5.4.4)$$

Using (5.4.3) and (5.4.4) in (5.4.2) and simplifying further we get

$$X''(z) = x_0'' \left[\tilde{U}''(z,0) - I \right] \tilde{A}''(z,0) \left[zI - \tilde{A}''(z,0) \right]^{-1} \quad (5.4.5)$$

$$X''(z) \left[zI - \tilde{A}''(z,0) \right] = \mathcal{U}(z) \quad (5.4.6)$$

where

$$\mathcal{U}(z) = x_0'' \left[\tilde{U}''(z,0) - I \right] \mathcal{P}''(z,D) \quad (5.4.7)$$

The moments of the queue lengths at Q2 can now be computed by differentiating (5.4.6) w.r.t. z successively. To obtain compact expression, we use the following notation. The superscripts (") are omitted, the n th moments of $X''(z)$, $\tilde{A}''(z,0)$ and $\mathcal{U}(z)$ w.r.t. z evaluated at $z=1$ are denoted as $X^{(n)}$, $A^{(n)}$ and $\mathcal{U}^{(n)}$ respectively. $X''(z)$, $\tilde{A}''(z,0)$ and $\mathcal{U}(z)$ evaluated at $z=1$ are denoted as X , A and \mathcal{U} respectively. With this notation, rewriting (5.4.6) and differentiating w.r.t. z we get

$$X(z)[zI - A(z)] = \mathcal{U}(z) \quad (5.4.8)$$

$$X^{(1)}(z)[zI - A(z)] + X(z)[I - A^{(1)}(z)] = \mathcal{U}^{(1)}(z) \quad (5.4.9)$$

$$X^{(2)}(z)[zI - A(z)] + 2X^{(1)}(z)[I - A^{(1)}(z)] - X(z)A^{(2)}(z) = \mathcal{U}^{(2)}(z) \quad (5.4.10)$$

$$\begin{aligned} X^{(3)}(z)[zI - A(z)] + 3X^{(2)}(z)[I - A^{(1)}(z)] \\ - 3X^{(1)}(z)A^{(2)}(z) - X(z)A^{(3)}(z) = \mathcal{U}^{(3)}(z) \end{aligned} \quad (5.4.11)$$

Multiplying (5.4.9)-(5.4.11) by e and evaluating at $z=1$ we get

$$X[I - A^{(1)}]e = \mathcal{U}^{(1)}e \quad (5.4.12)$$

$$2X^{(1)}[I - A^{(1)}]e = X A^{(2)}(z)e + \mathcal{U}^{(2)}e \quad (5.4.13)$$

$$3X^{(2)}[I - A^{(1)}]e = 3X^{(1)}A^{(2)}e + XA^{(3)}e + U^{(3)}e \quad (5.4)$$

Let π be the stationary vector of A . Then it can be verified that $[I - A + e\pi]$ is nonsingular and hence we get

$$\pi[I - A + e\pi] = \pi = \pi[I - A + e\pi]^{-1} \quad (5.4.1)$$

Adding both sides of (5.4.9) by $X^{(1)}e\pi$ and using (5.4.15) we get

$$X^{(1)} = X^{(1)}e\pi + [U^{(1)} - X(I - A^{(1)})][I - A + e\pi]^{-1} \quad (5.4.16)$$

Substituting (5.4.16) in (5.4.13) and using (5.4.12) we get

$$X^{(1)}e = \frac{1}{2(1-\rho)} \left\{ XA^{(2)}(z)e + U^{(2)}e + 2[U^{(1)} - X(I - A^{(1)})][I - A + e\pi]^{-1}\beta \right\} \quad (5.4.17)$$

where $\beta = A^{(1)}e$ and $\rho = \pi\beta$

Multiplying both sides of (5.4.10) by $X^{(2)}e\pi$ and using (5.4.15) we get

$$X^{(2)} = X^{(2)}e\pi + \left\{ U^{(2)}(z) - 2X^{(1)}[I - A^{(1)}] + XA^{(2)} \right\} [I - A + e\pi]^{-1} \quad (5.4.18)$$

Substituting (5.4.18) in (5.4.14) and using (5.4.12) we get

$$X^{(2)}e = \frac{1}{3(1-\rho)} \left\{ 3X^{(1)}A^{(2)}e + XA^{(3)}e + U^{(3)}e + 3 \left[XA^{(2)} + U^{(2)} - 2X^{(1)}[I - A^{(1)}] \right] [I - A + e\pi]^{-1}\beta \right\} \quad (5.4.19)$$

Hence $A^{(n)}$ and $U^{(n)}$ for $n=0,1,2$ and 3 should be known for the computation of the first two moments and their evaluation is considered next

Using (3.6.5) it can be verified that $A^{(n)}$ is given by

$$A^{(n)} = \mathcal{P}^{(n)} = \left. \left(\frac{d}{dz} \right)^n \mathcal{P}(Z,D) \right|_{z=1} = \sum_{k=0}^{\infty} \frac{k!}{(k-n)!} P(k,D) \quad (5.4.20)$$

$\mathcal{P}^{(n)}$, the n^{th} moment of $\mathcal{P}(z,t)$ has to be computed using numerical integration for a general service time distribution and the details are given in Sec 3.3 of Neuts[1]. However, since the matrices $P(k,D)$ are evaluated using the procedure given in Sec 4.4, computation of $A^{(n)}$ is straight forward and requires only moderate computational effort

Let the n^{th} moment of $\tilde{U}''(z,0)$ be denoted as $U^{(n)}$. Then using (5.4.7) we get

$$\mathcal{U}^{(n)} = x_0'' \sum_{m=0}^n \mathcal{U}^{(m)} \mathcal{P}^{(n-m)} - x_0'' \mathcal{P}^{(n)} \quad (5.4.21)$$

Finally we compute $\mathcal{U}^{(n)}$ for $n=0-3$ using (3.6.11) and are given by

$$\mathcal{U}^{(1)} = [-R''(0)]^{-1} \left\{ \Lambda'' p_0'' + (1-p_0'')(\mathcal{J}_0 + \mathcal{J}_1) \right\} \quad (5.4.22)$$

$$\mathcal{U}^{(2)} = [-R''(0)]^{-1} \left\{ (1-p_0'')(\mathcal{J}_0 + 2\mathcal{J}_1 + \mathcal{J}_2) \right\} \quad (5.4.23)$$

$$\mathcal{U}^{(3)} = [-R''(0)]^{-1} \left\{ (1-p_0'')(\mathcal{J}_0 + 3\mathcal{J}_1 + 3\mathcal{J}_2 + \mathcal{J}_3) \right\} \quad (5.4.24)$$

where

$$\mathcal{J}_n = \Lambda'' \frac{1}{D} \int_{w=0}^D \mathcal{P}^{(n)}(z, w) dw \Big|_{z=1} = \Lambda'' \frac{1}{D} \int_{w=0}^D dw \sum_{k=0}^{\infty} \frac{k!}{(k-n)!} P''(k, w) \quad (5.4.25)$$

\mathcal{J}_n for $n=0$ can be evaluated using (4.5.9), and for $n \geq 1$ they have to be numerically integrated using (5.4.25). This completes the set of equations required for the evaluation of the first two moments of the queue lengths of Q2. The higher order moments can also be computed along the same lines.

Next we consider the evaluation of the corresponding moments for Q1. It can be noted that the equations (5.4.1)-(5.4.19) are also valid for Q1 if the parameters of Q2 are replaced by those of Q1. Hence we need to consider only the details on the computation of the moments $\mathcal{A}^{(n)}$ and $\mathcal{U}^{(n)}$ corresponding to Q1.

Using (3.6.13) it can be verified that, for Q1, $\mathcal{A}^{(n)}$ is given by

$$\begin{aligned} \mathcal{A}^{(n)} &= \sum_{\ell=1}^{\infty} C(\ell-1) \mathcal{P}^{(n)} = \sum_{\ell=1}^{\infty} C(\ell-1) \left(\frac{d}{dz} \right)^n \mathcal{P}(z, \ell D) \Big|_{z=1} \\ &= \sum_{\ell=1}^{\infty} C(\ell-1) \sum_{k=0}^{\infty} \frac{k!}{(k-n)!} P'(k, \ell D) \end{aligned} \quad (5.4.26)$$

Let the n^{th} moment of $\tilde{U}'(z, 0)$ be denoted as $\tilde{U}^{(n)}$. Then using (5.4.7) we get

$$\mathcal{U}^{(n)} = x_0' \sum_{m=0}^n \tilde{U}^{(m)} \mathcal{P}^{(n-m)} - x_0' \mathcal{P}^{(n)} \quad (5.4.27)$$

Finally we compute $\tilde{U}^{(n)}$ for $n=0-3$ using (3.6.14). These are given by

$$\tilde{U}^{(1)} = [-R'(0)]^{-1} \left\{ \Lambda' p'_0 + (1-p'_0)(\mathcal{J}'_0 \mathcal{J}_0 + \mathcal{J}'_1 \mathcal{J}_0 + \mathcal{J}_0 \mathcal{J}'_1) \right\} \quad (5.4.28)$$

$$\tilde{U}^{(2)} = [-R'(0)]^{-1} \left\{ (1-p'_0) [2(\mathcal{J}'_1 \mathcal{J}_1 + \mathcal{J}'_1 \mathcal{J}_0 + \mathcal{J}'_0 \mathcal{J}_1) + \mathcal{J}'_0 \mathcal{J}_2 + \mathcal{J}'_2 \mathcal{J}_0] \right\} \quad (5.4.29)$$

$$\begin{aligned} \tilde{U}^{(3)} = [-R'(0)]^{-1} \left\{ (1-p'_0) [3(\mathcal{J}'_2 \mathcal{J}_0 + \mathcal{J}'_1 \mathcal{J}_1 + \mathcal{J}'_0 \mathcal{J}_2 \right. \\ \left. + \mathcal{J}'_2 \mathcal{J}_1 + \mathcal{J}'_1 \mathcal{J}_2) + \mathcal{J}'_0 \mathcal{J}_3 + \mathcal{J}'_3 \mathcal{J}_0] \right\} \end{aligned} \quad (5.4.30)$$

where

$$\mathcal{J}'_n = \underline{\Lambda}' \frac{1}{D} \int_{w=0}^D \mathcal{P}^{(n)}(z, w) dw \Big|_{z=1} = \underline{\Lambda}' \frac{1}{D} \int_{w=0}^D dw \sum_{k=0}^{\infty} \frac{k!}{(k-n)!} P'(k, w) \quad (5.4.31)$$

$$\mathcal{J}_n = \sum_{\iota=0}^{\infty} F(\iota) \sum_{k=0}^{\infty} \frac{k!}{(k-n)!} P'(k, \iota D) \quad (5.4.32)$$

$P'(k, \iota D)$ can be computed using the procedure given in Sec. 4.4. Hence the first two moments of Q1 are also in a computable form. However, compared to Q2, computational complexity is increased significantly for Q1, due to the presence of the terms \mathcal{J}'_n 's in the expression for the moments of $\tilde{U}'(z, 0)$.

5.5. COMPUTATION OF Y'_0 AND Y''_0

In this section, we consider the evaluation of the stationary probability $y''(0, j)$ of $Q''(\cdot)$ visiting the state $(0, j)$ at an arbitrary time instant and the corresponding probability $y'(0, j)$ of $Q'(\cdot)$. Let y'_0 and y''_0 denote the $1 \times MN$ vectors whose j^{th} elements are $y'(0, j)$ and $y''(0, j)$ respectively. The row sums of these vectors give the probabilities p'_0 and p''_0 required for the evaluation of the QLDs. The approach used here is analogous to that used in [3], [4] for computing y_0 of the N/G/1 and MMPP/G/1 queues with FCFS discipline.

We concentrate on Q2 first. Let $\Phi_{m\ell}^{1j'}(t)$ denote the expected number of visits of the semi-Markov process $Q''(\cdot)$ to the state (m, ℓ) in the interval $(0, t]$ given that it started from the state $(1, j')$ at time 0. First we obtain

an expression for $y''(0,j,t)$, the conditional probability of the semi-Markov process visiting the state $(0,j)$ at time t given that at time 0 the state was $(1,j')$ i.e

$$y''(0,j,t) \equiv P [X''(t)=0, \underline{j}(t) = j \mid X''(0) = 1, \underline{j}(0) = j'] \quad (5.5.1)$$

Let σ be the instant before time t when Q2 became empty for the last time. Hence in the interval $(0,\sigma]$ $Q''(\cdot)$ goes from the level i to level 0 (i.e the set of states $\{(0,\ell), 1 \leq \ell \leq MN\}$) and in the interval (σ,t) it continues to be in level 0 (in other words there are zero arrivals at Q2 in (σ,t)).

It may be noted that the probability that the state of the MRP at time σ is equal to $(0,j)$ given that it was in state $(1,j')$ at time 0 is also equal to the average number of visits of the MRP to the state $(0,j)$ at time σ given that the MRP was in state $(1,j')$ at time 0. Using this and considering the above two chains of events we get

$$y''(0,j,t) = \sum_{\ell=1}^{MN} \int_{\sigma=0}^t d\Phi_{0\ell}''(\sigma) P_{\ell j}''(0,t-\sigma) \quad (5.5.2)$$

Removing the dependence on t by applying the Key renewal theorem Wolff[5], Medhi[6] we get

$$y''(0,j) = \lim_{t \rightarrow \infty} y''(0,j,t) = \sum_{\ell=1}^{MN} \frac{1}{m''(0,\ell)} \int_{t=0}^{\infty} P_{\ell j}''(0,t) dt \quad (5.5.3)$$

where $m''(0,\ell)$ is the mean recurrence time (MRT) of the state $(0,\ell)$ in $Q''(\cdot)$. Analogous to that in Sec 5.2, it may be verified that the MRT of $(0,\ell)$ in $Q''(\cdot)$ is also the MRT of MRP $\tilde{K}_0''(1,s)$. Applying Theorem (2.11) of Hunter[2] we get

$$m''(0,\ell) = (k_0'' \hat{\mu}'')[(k_0'')_{\ell}]^{-1} \quad (5.5.4)$$

where the i^{th} element of $\hat{\mu}''$ gives the average length of the busy cycle starting in phase i of MMPP \underline{z} .

Using (5.5.4) in (5.5.3) and writing it in vector form, we get

$$y_0'' = (k_0'' \hat{\mu}'')^{-1} k_0'' \int_{t=0}^{\infty} P''(0,t) dt \quad (5.5.5)$$

$$= (k_0'' \hat{\mu}'')^{-1} k_0'' [-R''(0)]^{-1} \quad (5.5.6)$$

Using (4.6.2), (4.6.4) and (4.5.21) in (5.5.6) we get

$$y_0'' = \left[k_0'' (\mu_1'' + \mu_2'') \right]^{-1} k_0'' [-R''(0)]^{-1} \quad (5.5.7)$$

$$= \left\{ (k_0'' \tilde{\mu}'') \left[\mu^{(1)''} + (k_0'' \tilde{\mu}'')^{-1} k_0'' \left[[-R''(0)]^{-1} e + (1-p_0') \frac{D}{2} e \right] \right] \right\}^{-1} k_0'' [-R''(0)]^{-1} \quad (5.5.8)$$

$$= \left[\mu^{(1)''} + x_0'' \left[[-R''(0)]^{-1} e + (1-p_0') \frac{D}{2} e \right] \right]^{-1} x_0'' [-R''(0)]^{-1} \quad (5.5.9)$$

$$p_0'' = y_0'' e \quad (5.5.10)$$

Next, we consider an interpretation for the term inside first the square bracket of (5.5.9). Let us consider the computation of $m''(1,j)$, the mean recurrence time of the state $(1,j)$ of the semi-Markov process $Q''(\cdot)$. Let δ_T denote the mean sojourn time of the process $Q''(\cdot)$ averaged over all possible states. Let ℓ_{1j} denote the MRT of the state $(1,j)$ of the SMC $Q''(\omega)$. It may be noted that ℓ_{1j} is also equal to the inverse of $x(1,j)$, the probability of the state $(1,j)$ of $Q''(\omega)$. By theorem 2.11 of Hunter [2], we get

$$m''(1,j) = \delta_T \ell_{1j} = \delta_T [x(1,j)]^{-1} \quad (5.5.11)$$

δ_T can be computed as follows. Let $\delta(1,j)$ denote the sojourn time of $Q''(\cdot)$ in the state $(1,j)$. Let δ_1 denote the $MN \times 1$ vector whose j^{th} element is $\delta(1,j)$. $\delta(0,j)$ can be computed as follows. Let the state $(0,j)$ be visited at a departure instant. Let t be the time when the next departure occurs. Given that the previous departure left the queue empty and the MMPP $\underline{2}$ in phase j , the probability that the next departure occurs at time t and leaves behind k

customers is given by $\sum_{j=1}^{MN} dB_k''(t)$ where j is the phase of MMPP $\underline{2}$ at time t

Hence the mean sojourn time of the state $(0, j)$ is given by-

$$\delta(0, j) = \sum_{k=0}^{\infty} \int_0^{\infty} t \sum_{j=1}^{MN} dB_k''(t) \quad (5.5.12)$$

Writing (5.5.12) in vector form we get

$$\delta_0 = \sum_{k=0}^{\infty} \int_0^{\infty} t dB_k''(t) e \quad (5.5.13)$$

The integral of (5.5.13) can be rewritten as-

$$\delta_0 = - \frac{d}{ds} \sum_{k=0}^{\infty} \int_0^{\infty} dB_k''(t) e^{-st} e \Big|_{s=0} = - \frac{d}{ds} \sum_{k=0}^{\infty} \tilde{B}_k''(s) e \Big|_{s=0} \quad (5.5.14)$$

Using (3.3.18) in (5.5.14) we get

$$\delta_0 = - \frac{d}{ds} \sum_{k=0}^{\infty} \sum_{m=1}^{k+1} \tilde{U}_m''(s) P''(k-m+1, D) e^{-sD} e \Big|_{s=0} \quad (5.5.15)$$

Expanding the summations of (5.5.15) and combining the like terms we get

$$\delta_0 = - \frac{d}{ds} \tilde{U}''(1, s) \mathcal{P}''(1, D) e^{-sD} e \Big|_{s=0} \quad (5.5.16)$$

where $\tilde{U}''(z, s)$ and $\mathcal{P}''(z, D)$ are the z transforms of $\tilde{U}_k''(s)$ and $P''(k, D)$ respectively. Differentiating (5.5.16) w.r.t s and using (4.6.4) and (4.5.14) and noting that $\tilde{U}''(1, 0)$ and $\mathcal{P}''(1, D)$ are stochastic we get

$$\delta_0 = - \frac{d}{ds} \tilde{U}''(1, s) \mathcal{P}''(1, D) e - \tilde{U}''(1, 0) \frac{d}{ds} \mathcal{P}''(1, D) e^{-sD} e \quad (5.5.17)$$

$$= -[R''(0)]^{-1} e + (1 - p_0') \frac{D}{2} e + \mu^{(1)''} e \quad (5.5.18)$$

Similarly δ_1 for $i > 0$ is obtained as follows

$$\delta_1 = \sum_{k=0}^{\infty} \int_0^{\infty} t dA_k''(t) e \quad (5.5.19)$$

$$= - \frac{d}{ds} \sum_{k=0}^{\infty} \int_0^{\infty} dA_k''(t) e^{-st} e \Big|_{s=0} = - \frac{d}{ds} \sum_{k=0}^{\infty} \tilde{A}_k''(s) e \Big|_{s=0} \quad (5.5.20)$$

$$= - \frac{d}{ds} \tilde{A}''(1, s) e \Big|_{s=0} \quad (5.5.21)$$

$$= \mu^{(1)''} e \quad (5.5.22)$$

The mean sojourn time of $Q''(\cdot)$ averaged all possible states is obtained using (5.5.18) and (5.5.22) and is given by

$$\delta_T = \sum_{i=0}^{\infty} x_1'' \delta_1 \quad (5.5.23)$$

$$= \mu^{(1)''} + x_0'' \left[[-R''(0)]^{-1} e + (1-p_0') \frac{D}{2} e \right] \quad (5.5.24)$$

It can now be noted that the term inside the first the square bracket of (5.5.9) actually denotes the mean sojourn time of the semi-Markov process $Q''(\cdot)$. Let the inverse of the mean sojourn time of $Q''(\cdot)$ be denoted as ξ^{**}

$$\xi^{**} = \frac{1}{\delta_T} = \left\{ \mu^{(1)''} + x_0'' \left[[-R''(0)]^{-1} e + (1-p_0') \frac{D}{2} e \right] \right\}^{-1} \quad (5.5.25)$$

Using (5.5.25), (5.5.11) and (5.5.9) can be rewritten as

$$\frac{1}{m''(1, j)} = \xi^{**} x''(1, j) \quad (5.5.26)$$

$$y_0'' = \xi^{**} x_0'' [-R''(0)]^{-1} \quad (5.5.27)$$

It is shown in Ramaswami [3] that in an N/G/1 queue the inverse of the mean sojourn time of the semi-Markov process $Q(\cdot)$ corresponding to the N/G/1 queue is also equal to the ratio of the expected number of renewals during a renewal interval of the input arrival process to the expected length of that renewal interval. Comparing (5.5.25) with Lemma 3.3.1 of [3], it can be concluded that this relationship is not valid for the priority system considered in this thesis.

The corresponding expression for y_0' can be obtained by replacing the

parameters of Q2 by those of Q1 in (5.5.7). Using (4.6.5) and (4.5.21) in the resulting expression we get

$$y'_0 = \left[k'_0 (\mu'_1 + \mu'_2) \right]^{-1} k'_0 [-R'(0)]^{-1} \quad (5.5.28)$$

$$= \left\{ (k'_0 \tilde{\mu}') \left[\mu^{(1)'} + (k'_0 \tilde{\mu}')^{-1} k'_0 \left[[-R'(0)]^{-1} e + (1-p''_0) \left([-R'(0)]^{-1} \Lambda' V(D) \sum_{\iota=0}^{\infty} F(\iota) \iota e + \frac{D}{2} e \right) \right] \right] \right\}^{-1} k'_0 [-R'(0)]^{-1} \quad (5.5.29)$$

$$= \left[\mu^{(1)'} + x'_0 \left([-R'(0)]^{-1} e + (1-p''_0) \left([-R'(0)]^{-1} \Lambda' V(D) \sum_{\iota=0}^{\infty} F(\iota) \iota e + \frac{D}{2} e \right) \right) \right]^{-1} x'_0 [-R'(0)]^{-1} \quad (5.5.30)$$

$$p'_0 = y'_0 e \quad (5.5.31)$$

It may be noted that y'_0 as well as the mean recurrence time of the state $(1, j)$ of the semi-Markov process $Q'(\cdot)$ can be expressed in terms of the mean sojourn time of $Q'(\cdot)$ averaged over all possible states. Let $m'(1, j)$ denote the MRT of the state $(1, j)$ of $Q'(\cdot)$ and let ξ^{*} denote the inverse of the mean sojourn time of $Q'(\cdot)$. Proceeding along the same lines as for Q2 it can be shown that

$$\frac{1}{m'(1, j)} = \xi^{*} x'(1, j) \quad (5.5.32)$$

$$y'_0 = \xi^{*} x'_0 [-R'(0)]^{-1} \quad (5.5.33)$$

$$\begin{aligned} \xi^{*} &= \left[\mu^{(1)'} + x'_0 \left([-R'(0)]^{-1} e + (1-p''_0) \left([-R'(0)]^{-1} \Lambda' V(D) \sum_{\iota=0}^{\infty} F(\iota) \iota e + \frac{D}{2} e \right) \right) \right]^{-1} \end{aligned} \quad (5.5.34)$$

It may be noted that the queue length densities of Q2 and Q1, at an arbitrary time instant, can also be computed along the same lines as in [1], [3]. As we

do not require these probabilities for our present work, we have not discussed them in detail

5.6. STATIONARY QUEUE LENGTH DISTRIBUTION OF Q1 AND Q2

The evaluation of the queue length distribution of Q1 and Q2 at their respective departure instants, is considered in this section. At the outset, the two basic issues involved in the evaluation, ie iteration and truncation are discussed. As noted in Sec 3.1, the queue length densities (QLDs) of Q1 and Q2, have to be computed iteratively as the matrices $Q''()$ and $Q'()$ are coupled. The dependence of the QLD of Q2 on the traffic offered at Q1 comes through the term " p_0' ". On the other hand, for the evaluation of the QLD of Q1, in addition to p_0'' , the QLD of Q2 must be known. In view of the lighter dependence of Q2 on traffic offered at Q1, Q2 turns out to be the obvious choice for starting the iteration. Assuming an initial value of p_0' , QLD of Q2 and p_0'' can be found. These in turn can be used to find the QLD and p_0' of Q1. This procedure can be repeated until p_0'' and p_0' stabilize.

For the computation of the QLDs using numerical methods, the infinite dimensional matrices $Q''(\infty)$ and $Q'(\infty)$ have to be truncated suitably. Let L' and L'' denote the dimensions of $Q''(\infty)$ and $Q'(\infty)$ after truncation. One standard way of determining these values is to make use of the " $\mu(\text{mean}) + 3\sigma(\text{standard deviation})$ " rule as in [1], [3]. The expression for the first two moments of the queue lengths of Q1 and Q2, developed in Sec 5.4 comes in handy for this purpose. Compared to the FCFS queue, in the prioritized queue, the evaluation of these moments demand significant computational effort. In view of the coupling between the queues, their values also keep changing with each iteration. Hence this method seems to be attractive if the storage and computational resources have to be minimized at the expense of implementation.

complexity Alternately, one may find the moments corresponding to traffic at Q2 with the FCFS discipline and choose L'' to be greater than this Similarly finding the moments corresponding to the total traffic offered at Q1 and Q2, L' may be chosen (The moments corresponding to FCFS discipline can be found using the results of Sec 5.4 by substituting either p'_0 or p''_0 to be equal to 1) The truncation index chosen using either method may have to be incremented by fixed amounts, if the tail probabilities computed thereby are not sufficiently small

Having chosen the truncation indices, at each stage of iteration, the QLD of either Q1 or Q2 has to be computed using numerical methods Since the evaluation procedure is the same for both Q1 and Q2, we shall concentrate only on the evaluation of the QLD of Q2 Let the truncation index for $Q''(\infty)$ be L We shall drop the superscript of (∞) for ease of notation For the evaluation of the QLD, we consider three methods (i) Gaussian elimination (ii) Block Toeplitz matrix inversion and (iii) Recursive method

The methods (i) and (ii) do not require the computation of x_0 using (5.2.2) and (5.2.3) The matrices G' and G'' are also not required Hence the computational as well as implementational efforts are less for these methods However, the storage requirements are of the order $O(n^2)$ and hence these methods are to be preferred only if the truncation indices are not large (less than 500) or equally if the traffic offered to Q1 and Q2 is not close to the capacity of the server

An alternative is to make use of the iterative schemes Kreiger[7], Stewart[8] Even though, their storage requirements are relatively small, at higher traffic rates, their convergence is slow In spite of this, until very recently, iterative schemes like the Block Gauss-Seidel iteration scheme [7], [8], [1] have been employed to obtain the QLDs Recently, Ramaswami [9] came

up with a recursive scheme which is considered to be a major breakthrough (see for e.g. Lucantonio[10]) It drastically reduces the storage requirements as well as the computational requirements In a typical example considered in [9], the computation time required for this scheme was found to be reduced by a factor of more than 1800 compared to the Block Gauss-Seidel procedure Further, while both x_0 and x_1 are required for the Gauss-Seidel procedure, only the former is required for the recursive scheme of [9] Unlike the methods (i) and (ii), here x_0 's and G 's have to be evaluated first

GAUSSIAN ELIMINATION METHOD

The Gaussian elimination procedure or its variants may be used to solve the truncated system of equations given by (5.6.1) A survey of the various methods under this category as well as the time and storage requirements of these methods are discussed in Kreiger[7] We consider one procedure in detail Let the truncated transition probability matrix of Q_2 be denoted as \hat{Q} The QLDs x_i for $i=0,1,2 \dots L-1$ can be evaluated by solving the set of equations given by

$$\hat{Q}x = x \quad x[\hat{Q} - I] = 0 \quad (5.6.1)$$

$$x_e = 1 \quad (5.6.2)$$

$$\hat{Q} = \begin{bmatrix} B_0 & B_1 & B_2 & & B_{L-2} & \sum_{m=L-1}^{\infty} B_m \\ A_0 & A_1 & A_2 & & A_{L-2} & \sum_{m=L-1}^{\infty} A_m \\ 0 & A_0 & A_1 & & A_{L-3} & \sum_{m=L-2}^{\infty} A_m \\ & & & & & \\ 0 & 0 & 0 & & A_0 & \sum_{m=1}^{\infty} A_m \end{bmatrix} \quad (5.6.3)$$

where $x = [x_0 \ x_1 \ x_2 \ \dots \ x_{L-1}]$, $x_1 = [x_{11} \ x_{12} \ \dots \ x_{1MN}]$, I is the $MNL \times MNL$

identity matrix, $\underline{e} = [e^T e^T e^T e^T e^T]^T$ and $\underline{0} = [0 \ 0 \ 0 \ 0]_{1 \times MNL}$. The elements in the last block column of \hat{Q} are chosen such that the row sums are equal to 1 as shown in (5.6.3). The normalization condition of eqn (5.6.2) can be included in the set of equations given by (5.6.1) by replacing the last column of \hat{Q} by the $MNL \times 1$ vector $[1 \ 1 \ 1 \ 1 \ 0]^T$. Let the modified matrix be denoted as \underline{Q} . The null row vector in the right hand side of (5.6.1) is also modified accordingly so that the last element in the row alone is changed to be 1. Let $\mathcal{R} \equiv \underline{Q} - \underline{I}$. With these modifications (5.6.1) becomes

$$\mathbf{x}\mathcal{R} = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1]_{1 \times MNL} \quad (5.6.4)$$

$$\mathcal{R} = \begin{bmatrix} F_0 & F_1 & F_2 & & F_{L-2} & F_{L-1} \\ E_0 & E_1 & E_2 & & E_{L-2} & G_{N-1} \\ 0 & E_0 & E_1 & & E_{L-3} & G_{N-2} \\ & & & & & \\ 0 & 0 & 0 & & E_0 & G_1 \end{bmatrix} \quad (5.6.5)$$

It may be noted that \mathcal{R} given in (5.6.5) is very nearly upper triangular. Using Gaussian elimination it can be made perfectly upper triangular and the QLDs can be found using "back substitution". We discuss this method in detail as it has the advantage of ease of implementation. Let the $(i,j)^{th}$ elements of \mathcal{R} be denoted as $\mathcal{R}(i,j)$. \mathcal{R} can be converted to be perfectly upper triangular by applying the row transformations on rows 2 to $MNL - 1$ repeatedly as follows

For $n=1, MNL-1$

For $i=1, MN-1$

For $j=1, MN$

$$\mathcal{R}(n+1, j) = \mathcal{R}(n+1, j) - \frac{\mathcal{R}(n, j)}{\mathcal{R}(n, n)} \mathcal{R}(n+1, j) \quad (5.6.6)$$

The elements in the RHS of (5.6.4) do not get affected by these transformations as the first $MN-1$ elements are zero. Hence x can be obtained using (5.6.4) with the modified matrix \mathcal{R} .

BLOCK TOEPLITZ MATRIX INVERSION METHOD

We next consider a more efficient method for the evaluation of the QLD based on Block Toeplitz Matrix Inversion. Using (5.6.4), the eigenvector x is given by

$$x = [0 \ 0 \ 0 \ 0 \ 1] \mathcal{R}^{-1} \quad (5.6.7)$$

The computation of the inverse of \mathcal{R} can be reduced to the computation of inverse of $MN \times MN$ matrices and their products as discussed in Blondia[11]. We consider this next. Let I , O be the $MN \times MN$ identity and the null matrix respectively. Let \mathcal{P} be the permutation matrix given by (5.6.9). Premultiplying \mathcal{R} by \mathcal{P} we get

$$\mathcal{P}\mathcal{R} = \left[\begin{array}{cccc|cc} E_0 & E_1 & E_2 & & E_{L-2} & G_{L-1} \\ & 0 & E_0 & E_1 & & E_{L-3} & G_{L-2} \\ & & 0 & 0 & E_0 & & E_{L-4} & G_{L-3} \\ & & & 0 & 0 & 0 & E_0 & G_1 \\ F_0 & F & F_2 & & F_{L-2} & F_{L-1} \end{array} \right] = \begin{bmatrix} E & G \\ F & C \end{bmatrix} \quad (5.6.8)$$

$$\mathcal{P} = \begin{bmatrix} 0 & I & O & O & . & . & 0 & 0 \\ 0 & 0 & I & O & & & 0 & 0 \\ & & & . & . & & & \\ 0 & 0 & 0 & 0 & & 0 & I & \\ I & 0 & 0 & 0 & & 0 & 0 & 0 \end{bmatrix} \quad (5.6.9)$$

If the inverse of A_0 of (5.6.3) exists then E_0^{-1} also exists and hence E^{-1}

exists Using the Schur-Banachiewicz formula for the inverse of block matrices [12] in (5.6.8) we get

$$[\mathcal{P}\mathcal{R}]^{-1} = \begin{bmatrix} \mathbb{E}^{-1} + \mathcal{E}\Delta^{-1}\mathcal{F} & -\mathcal{E}\Delta^{-1} \\ -\Delta^{-1}\mathcal{F} & \Delta^{-1} \end{bmatrix} \quad (5.6.10)$$

where $\Delta = \mathbb{C} - \mathbb{F}[\mathbb{E}^{-1}]\mathbb{G}$, $\mathcal{E} = [\mathbb{E}^{-1}]\mathbb{G}$ and $\mathcal{F} = \mathbb{F}[\mathbb{E}^{-1}]$. The matrix Δ is known as the Schur complement of the matrix \mathbb{E} and is non-singular if \mathbb{E} is non-singular. Since the matrix Δ is an $MN \times MN$ matrix, the computation of Δ^{-1} does not demand much computational effort. Next an efficient procedure for the computation of \mathbb{E}^{-1} is obtained, exploiting the fact that \mathbb{E} is a block Toeplitz matrix.

It may be recalled that a matrix T whose $(i,j)^{\text{th}}$ element, T_{ij} , is a function of $(i-j)$ is generally called as Toeplitz matrix [13]. When T_{ij} itself is a matrix, T is called a block Toeplitz matrix. It may be noted that the Matrix \mathbb{E} is a block upper triangular Toeplitz matrix (BUTT). We next state some of the properties of BUTT matrices (see for e.g. Trench[14], [15], Jain[16]).

- (i) The inverse of a BUTT matrix is also a BUTT matrix.
- (ii) The elements in the top most $m \times m$ block of the inverse of a $(m+1, m+1)$ BUTT matrix is the same as the inverse of the $m \times m$ BUTT matrix obtained by leaving the last row and last column of the $(m+1, m+1)$ matrix. Let us assume that the inverse of the $m \times m$ BUTT matrix is known; for finding the inverse of the $(m+1, m+1)$ BUTT matrix only the $(1, m+1)^{\text{th}}$ element of the inverse matrix needs to be found.

Let us consider an example where $m=2$. Using the above property the elements of the inverse of a 3×3 BUTT matrix B can be related to that of a 2×2 BUTT matrix A as follows

$$A = \begin{bmatrix} R_0 & R_1 \\ 0 & R_0 \end{bmatrix}^{-1} = \begin{bmatrix} P_0 & P_1 \\ 0 & P_0 \end{bmatrix}$$

$$B = \begin{bmatrix} R_0 & R_1 & R_2 \\ 0 & R_0 & R_1 \\ 0 & 0 & R_0 \end{bmatrix}^{-1} = \begin{bmatrix} Q_0 & Q_1 & Q_2 \\ 0 & Q_0 & Q_1 \\ 0 & 0 & Q_0 \end{bmatrix} = \begin{bmatrix} P_0 & P_1 & Q_2 \\ 0 & P_0 & P_1 \\ 0 & 0 & P_0 \end{bmatrix}$$

Hence, if A is known, then for finding B only the matrix Q_2 needs to be evaluated

The inverse of a BUTT matrix can be found iteratively using the procedure for inversion of block Toeplitz matrices developed by Akaike[13]. The application of the properties (i) and (ii) of a BUTT matrix significantly reduces the computational effort. We next consider more details of this procedure.

A $(p+1)^{\text{th}}$ order BUTT matrix, L_{p+1} can be expressed in terms of the p^{th} order BUTT matrix L_p as follows

$$L_{p+1} = \begin{bmatrix} A_0 & \tilde{A}_p \\ 0_p & L_p \end{bmatrix}$$

where A_0 , \tilde{A}_p , 0_p and L_p are 1×1 , $1 \times p$, $p \times 1$ and $p \times p$ block matrices of size $d \times d$. $\tilde{A}_p = [A_1 \ A_2 \ A_3 \ \dots \ A_{p-1} \ A_p]$. It may be noted that $[A_0 \ \tilde{A}_p]$ correspond to the first $p+1$ elements in the top row of the matrix E to be inverted. 0_p is a null block matrix.

The symbol tilde (\sim) denotes the block transposition of matrices, i.e. interchange of the $(i,j)^{\text{th}}$ block with the $(j,i)^{\text{th}}$ block. Let the symbol $(\hat{\cdot})$ denote the reversal of the ordering of the blocks in a row, i.e.

$$\tilde{A}_p = [A_1 \ A_2 \ A_3 \ \dots \ A_{p-1} \ A_p] \Rightarrow \hat{\tilde{A}}_p = [A_p \ A_{p-1} \ \dots \ A_3 \ A_2 \ A_1]$$

The $(p+1)^{\text{th}}$ order BUTT inverse matrix, L_{p+1}^{-1} can be expressed in terms of the p^{th} order BUTT matrix M_p as follows

$$L_{p+1}^{-1} = \begin{bmatrix} Q_0 & \tilde{Q}_p \\ 0_p & M_p \end{bmatrix}$$

where Q_0 , \tilde{Q}_p , 0_p and M_p are 1×1 , $1 \times p$, $p \times 1$ and $p \times p$ block matrices of size $d \times d$.

It can be noted that if we know the elements in the first row of L_{p+1}^{-1} , M_p can

be obtained by cyclic shift of the first row. Using the procedure given in Akaike[13], the block matrix $\tilde{\mathcal{F}}_p$ can be iteratively computed and is given by

$$\tilde{\mathcal{F}}_{p+1} = \left\{ Q_0 \left[A_{p+1} + A_0 \sum_{n=1}^p \tilde{\mathcal{F}}_n A_n \right] Q_0 \quad \tilde{\mathcal{F}}_p \right\} \quad (5.6.11)$$

$$Q_0 = A_0^{-1}$$

RECURSIVE PROCEDURE

A recursive procedure for the computation of x_i , for $i \geq 1$ can be obtained using (5.4.1). However, this recursion suffers from "catastrophic cancellation" (see for e.g. Forsythe [17]) which results from subtracting small quantities of the same order. An alternative and numerically stable recursive procedure is suggested by Ramaswami[9]. Using this procedure, x_i can be computed as follows

$$x_i = \left[x_0 \bar{B}_i + \sum_{j=1}^{i-1} x_j \bar{A}_{i-j+1} \right] \left[I - \bar{A}_1 \right]^{-1} \quad (5.6.12)$$

$$\bar{B}_v = \sum_{l=v}^{\infty} B_l G^{l-v} \quad \bar{A}_v = \sum_{l=v}^{\infty} A_l G^{l-v}$$

where $G = G(z, s)$ evaluated at $z=1$ and $s=0$. Since all the quantities in this recursion are non-negative, it does not suffer from the catastrophic cancellation suffered by other recursions. Further, as observed in [9], the implementation of (5.6.12) can be done efficiently by noting that as $i \rightarrow \infty$, $\bar{B}_i, \bar{A}_i \rightarrow 0$. Hence choosing a large index L , \bar{B}_L, \bar{A}_L can be set to be the null matrices. The matrices \bar{B}_k, \bar{A}_k for k less than L is computed using the backward recursions

$$\bar{B}_k = B_k + \bar{B}_{k+1} G \quad \text{and} \quad \bar{A}_k = A_k + \bar{A}_{k+1} G \quad \text{for } 0 \leq k \leq L-1 \quad (5.6.13)$$

EVALUATION OF THE QLDs FOR SOME SPECIAL CASES

I Q1 FINITE SIZED AND Q2 INFINITE SIZED

When the traffic offered to Q1 is very high, the practical sizes used for Q1 may not be large enough to be treated to be infinite. Hence a finite sized Q1 and an infinite sized Q2 is an interesting case of some practical importance. The evaluation of the QLDs can be carried out along the same lines as far the infinite size case. Let L be the buffer size of Q1. The truncation index L' for Q1 becomes equal to $L+1$. Proceeding along the same lines as in Barlow [18] and Blondia [11], it can be shown that the expression for y'_0 for the finite sized Q1 is also given by (5.5.30)

II EVALUATION OF THE QLD OF Q2 AND THE MOMENTS OF THE QUEUE LENGTHS OF Q1

Again we consider the case where the traffic offered to Q1 is high. In this case the truncation indices required for Q1, with an infinite sized Q1, becomes high and hence the storage and computational complexity also become high. In this case, one may be contented with the knowledge of the first two moments of the queue lengths of Q1. This can be computed as follows. First, starting with an initial value of p'_0 , the QLD of Q2 is found. Then this is used to compute the new value of x'_0 and y'_0 using (5.2.3) and (5.5.30). These steps are repeated until p'_0 and p''_0 stabilize. Knowing the p''_0 and the QLD of Q2, the moments of the queue lengths of Q1 can be found using the results of Sec. 5.4

5.7. AN APPROXIMATE MODEL FOR THE TIME PRIORITY SYSTEM

We have so far considered in detail an exact model for the study of a non-preemptive MMPP/D/1 priority system. The computational complexity and the storage requirements increase by a factor of $O(n^2)$ with this model compared to

that for a queue with a simple FCFS discipline. This increase occurs as the inter-departure time of the cells from Q1 (IDT 1) depend on the phase of the MMPP 2 at the previous departure instant. In other words, IDT 1 is not independent and identically distributed (i.i.d) if the phase of the MMPP 2 is not taken into account. We next consider an approximate model given in [19] which approximates the IDT 1 to be i.i.d and treats it to be independent of the *actual phase* of MMPP 2. For this approximate model, numerical results can be obtained with significantly lower computational and storage requirements.

We consider the scenario in which this approximate model can become almost as "good" as the exact model. When the traffic offered at Q2 is low, the probability that busy periods of Q2 exceed the transmission times (D) of a few cells become negligibly small. In addition to that, let the composite traffic to Q2, originate from a large number of sources, then the burstiness of the composite traffic gets smoothened compared to that of the individual sources and the arrival rates of the MMPP 2 at various arrival phases become almost equal. Hence the busy period distribution of Q2 (BPD) conditioned on the phase of MMPP 2 at the beginning of the BP tends to be only marginally different with different phases. Let us consider a "hypothetical process" which mimics the doubly stochastic Poisson process by a singly stochastic process and captures the average behaviour of the former. One such process is the simple Poisson process whose arrival rate is chosen to be the weighted average of the arrival rates of MMPP 2. However, we can treat the hypothetical process to be more general than this by choosing its BPD to be the weighted average of the BPD of MMPP 2 corresponding to various possible starting phases. It may be noted that, when the transition rates of MMPP 2 are small compared to the cell transmission time, over a short period of time a MMPP behaves like a simple Poisson process. The hypothetical process may become a

good approximation even in this case

It should be pointed out at this point that the approximate model of [19] discussed next, was originally "thought" to be exact. However, it was pointed out by Ramaswami[20], that the model of [19] could at best be only a good approximation. It was observed in [20] that the exact model should keep track of the phase of MMPP 1 and MMPP 2 simultaneously. It should be acknowledged here, that the exact model of Sec 3.2 was born out of incorporating the suggestions in [20].

From now on, we shall denote the exact model and the approximate model as model I and II respectively. In order to reduce the set of symbols we shall use the same symbols for model II as those used in sections 3.2-5.6 to denote the various parameters pertaining to Q1 and Q2. However, to retain uniformity we denote the phase of MMPP 1 and MMPP 2 at τ'_n, τ''_n as J'_n and J''_n respectively, at an arbitrary time instant t they are denoted as $J'(t)$ and $J''(t)$.

As a consequence of the assumption of independence of IDT 1 on the *actual* phase of MMPP 2, the following simplifications result with model II

- 1 $\{ (X'_n, J'_n), \tau'_n - \tau'_{n-1} \geq 1 \}$ and $\{ (X''_n, J''_n), \tau''_n - \tau''_{n-1} \geq 1 \}$ form Semi-Markov Chains (SMCs) with the state space $\{0,1\} \times \{1,2, \dots, M\}$ and $\{0,1\} \times \{1,2, \dots, N\}$ respectively
- 2 $A''_m(t), B''_m(t)$ and $P''(m,t)$ become $N \times N$ matrices, $A'_m(t), B'_m(t)$ and $P'(m,t)$ become $M \times M$ matrices
- 3 $H''(t)$ and $H'(t)$, the c d f of IDT of cells from Q2 and Q1, become scalar functions. Let them be denoted as $H''(t)$ and $H'(t)$ respectively
- 4 The matrices $C(k)$ and $F(l)$ characterizing the distribution of the busy period and additional busy period of Q2 become scalars. Let them be denoted as $C(k)$ and $F(l)$ respectively

With model II, the matrices $B'_m(t), A'_m(t), B''_m(t)$ and $A''_m(t)$ are defined as

follows

$[A'_m(t)]_{1j} = P\{ \text{Given that a cell departed from Q1 at time 0, leaving at least one cell in Q1 and the arrival process MMPP 1 in phase 1, the next departure occurs at no later than time } t \text{ with MMPP 1 in phase } j, \text{ and in the intervening period there were } m \text{ arrivals} \}$

$[B'_m(t)]_{1j} = P\{ \text{Given that a cell departed from Q1 at time 0, leaving Q1 empty and the arrival process MMPP 1 in phase 1, the next departure occurs at no later than time } t \text{ with MMPP 1 in phase } j \text{ and leaves behind } m \text{ cells in Q1} \}$

$[A''_m(t)]_{1j} = P\{ \text{Given that a cell departed from Q2 at time 0, leaving at least one cell in Q2 and the arrival process MMPP 2 in phase 1, the next departure occurs at no later than time } t \text{ with MMPP 2 in phase } j, \text{ and in the intervening period there were } m \text{ arrivals} \}$

$[B''_m(t)]_{1j} = P\{ \text{Given that a cell departed from Q2 at time 0, leaving Q2 empty and the arrival process MMPP 2 in phase 1, the next departure occurs at no later than time } t \text{ with MMPP 2 in phase } j, \text{ and leaves behind } m \text{ cells in Q2} \}$

Comparing these with those defined in Sec 3.3, it can be noted that we have replaced the terms MMPP 1 and MMPP 2 with the terms MMPP 1 and MMPP 2 respectively. In fact, we have written the expressions in Secs 3.1-5.6 such that if we replace the parameters of MMPP 1 by those of MMPP 1 and those of MMPP 2 by MMPP 2 we get the various expressions corresponding to model II. Analogous to that in Sec 3.3, the $(i,j)^{th}$ elements of $P'(m,t)$, $U'_k(t)$, $P''(m,t)$ and $U''_k(t)$ are defined as follows

$[P'(n,t)]_{1,j}$	$P[N'(t)=n, J'(t)=j \mid N'(0)=0, J'(0)=1]$
$N'(t)$	No of arrivals at Q1 from MMPP 1 in $(0,t]$
$[U'_k(t)]_{1,j}$	$P[\text{Busy period of Q1 starts at or before time } t,$ $N'(t)=k, J'(t)=j \mid X'(0)=0, J'(0)=1]$
$[P''(n,t)]_{1,j}$	$P[N''(t)=n, J''(t)=j \mid N''(0)=0, J''(0)=1]$
$N''(t)$	No of arrivals at Q2 from MMPP 2 in $(0,t]$
$[U''_k(t)]_{1,j}$	$P[\text{Busy period of Q2 starts at or before time } t,$ $N''(t)=k, J''(t)=j \mid X''(0)=0, J''(0)=1]$

Proceeding along the same lines as in Sec 3.4 and replacing the parameters of MMPP 2 by those of MMPP 2 we get

$$A_m''(t) = P''(m,D)u(t-D) \quad (5.7.1)$$

$$B_m''(t) = \sum_{k=1}^{m+1} U_k''(t-D)P''(m-k+1,D)u(t-D) \quad (5.7.2)$$

$$\frac{d}{dt}U_k''(t) = \frac{(1-p_0')}{D} \int_0^{Du(t-D_-)+tu(D-t)} P''(0,w)\Lambda'' dw P''(k-1,t-w) + p_0'\delta_{1k}P''(0,t)\Lambda'' \quad (5.7.3)$$

where p_0' is the probability that Q1 is empty at an arbitrary time instant and δ_{1k} is the Kronecker delta function

To deduce the corresponding expressions for Q1 we start with the expression for $\frac{dH}{dt}(t)$. It is non zero only for integral multiples of D sec and is equal to the probability that the busy period of Q2 (BP), which starts when Q1 is not empty, is $t-D$ for $t \geq D$ sec. Next we define $N \times N$ matrices $\underline{G}''^{(m)}(t)$ whose the $(i,j)^{th}$ entry denote the probability that the BP which starts with m cells and with MMPP 2 in phase 1 ends at or before time t with phase j . Let $d\underline{G}''^{(m)}(kD)$ be denoted as $\underline{G}_k''^{(m)}$ for $k=1,2$. We define $\underline{G}_0''^{(0)}$ to be the $N \times N$ identity matrix

Then $\underline{G}''^{(m)}(t)$ can be expressed as

$$\underline{G}^{(m)}(t) = \sum_{k=m}^{\infty} G_k^{(m)}(t) u(t-kD) \quad (5.7.4)$$

Analogous to that in Sec 3.5, we consider the following chain of conditional events

- (i) The phase of MMPP 2 is equal to 1 at the beginning of the service of a Q1 cell given Q2 is empty
- (ii) Busy period of Q2 starts with m customers with the MMPP 2 in phase ℓ given that MMPP 2 was in phase 1 at the beginning of the service for the Q1 cell
- (iii) The busy period is of duration kD sec and ends in phase j given that the busy period started with m cells with MMPP 2 in phase ℓ

Let $C(k)$ denote the probability that a BP which follows a Q1 service is of duration kD sec for k equal to 0, 1, 2, ... Using the above conditional events it can be shown that

$$C(k) = \frac{1}{P'_0} \sum_{m=(1-\delta_{0k})}^k y_0'' P''(m,D) G_k^{(m)} e \quad (5.7.5)$$

$$\frac{dH'(t)}{dt} = C(k-1) \delta'(t-kD) \quad (5.7.6)$$

where the 1th entry of the $1 \times N$ vector y_0'' gives the probability that Q2 is empty and the arrival phase of Q2 is 1 at an arbitrary time instant and e is an $N \times 1$ vector given by $[1 \ 1 \ \dots \ 1]^T$ $\delta'(t)$ is the Dirac delta function

Using (5.7.6) and (3.3.5), it can be shown that, for Q1

$$A'_n(t) = \sum_{k=1}^{\infty} u(t-kD) P'(n,kD) C(k) \quad (5.7.7)$$

Proceeding along the same lines as in Sec 3.3 we get

$$B'_m(t) = \sum_{k=1}^{m+1} U'_k(t-D) P'(m-k+1,D) u(t-D) \quad (5.7.8)$$

Replacing $F(\iota)$ by $F(\iota)I$ in (3.5.15) we get

$$\begin{aligned} \frac{dU_k'}{dt}(t) &= \sum_{\iota=0}^{\infty} \sum_{n=(k-1)\delta_{10}}^{k-1} F(\iota) \left\{ u(t-\iota D) \int_0^{Du(t-\iota+1D_-)+(t-\iota D)u(\iota+1D-t)} P'(0, t-\iota D-w) \Lambda' dw P'(n, w) \right. \\ &\quad \left. P'(k-n-1, \iota D) \right\} \frac{(1-p_0'')}{D} + p_0'' \delta_{1k} P'(0, t) \Lambda' \end{aligned} \quad (5.7.9)$$

Here, p_0'' is the probability of Q2 being empty at an arbitrary time instant

Let x_{ni}'' , the i^{th} element of the $1 \times N$ vector x_n'' , denote the probability that n cells are left in Q2 at a departure instant τ_k'' when the arrival phase is i . Then, the probability that $ABP = \iota D$ sec is given by

$$F(\iota) = \sum_{n=1}^{\iota} x_n'' G_{\iota}^{(n)} e \quad (5.7.10)$$

It can be noted that the matrices $G_{\iota}^{(n)}$ can be computed using the recursive procedure given in Sec. 4.3 by replacing the parameters of MMPP 2 by those of MMPP 2.

The expressions for y_0'' and y_0' can be obtained using the results of Secs. 4.5-5.5 by replacing the parameters of MMPP 1 and MMPP 2 by those of MMPP 1 and MMPP 2, respectively. The fact that $F(\iota)$ is a scalar results in the following simplified expression for μ_2'

$$\tilde{\mu}_2' = [-R'(0)]^{-1} e + (1-p_0'') \left\{ \frac{D}{2} e + \sum_{\iota=1}^{\infty} \iota D F(\iota) e \right\} \quad (5.7.11)$$

Finally, it can be noted that model II can also be used to compute the QLD corresponding to the case where either Q1 or Q2 is modelled as a Poisson process. In this case the arrival rates in all the phases of the corresponding MMPP is chosen to be equal. Further, model II can also be used to study the case when MMPP 2 is approximated by a Poisson process. This can be done by choosing the arrival rate in each phase of the MMPP to be the same and equal to that of the weighted average of the arrival rate corresponding to the

actual MMPP

5.8 SOME ASSUMPTIONS AND APPROXIMATIONS FOR NUMERICAL COMPUTATIONS

In this section the assumptions and approximations made for the numerical computation of the queue length densities (QLDs) at Q1 and Q2 are considered. First we consider the assumptions made. We assume the traffic to Q1 and Q2 to originate from on/off sources. It may be recalled that in Sec. 2.3 we noted that, using a generalized on/off source model, it is possible to model any arbitrary source reasonably accurately. For simplicity all the sources which generate the traffic to a particular queue (either Q1 or Q2) are assumed to have the same characteristics. The sources which generate the traffic to Q1 may in general be dissimilar to those generating the traffic to Q2. The sources are assumed to be characterized in terms of average on duration, percentage on duration and bit rate during on duration. We shall assume the sources to be CBR sources i.e. in the on period the cells arrive at periodic intervals and no cells arrive in the off period. The on and off durations are exponentially distributed. An output link capacity of 150 Mbps and a cell size of 53 bytes are also assumed. Next, we consider the approximations made.

COMPUTATION OF THE MODEL PARAMETERS OF MMPP

The traffic to Q1 and Q2 are assumed to be approximated by two 2 phase MMPPs using the method proposed in Heffes[4]. We considered the details of matching the moments of the composite traffic from a number of sources with that of a 2 phase MMPP in Sec. 2.4. The equations required for the evaluation of the parameters of the 2 phase MMPP has been obtained in Heffes [4] and are reproduced here for ease of reference. The arrival rate and the infinitesimal generator matrices are denoted as Λ and Q^* and their elements are given by

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad Q^* = \begin{bmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{bmatrix}$$

These elements are obtained by considering the moments of the interarrival time distribution of the renewal process characterizing the arrivals from a source at two short time intervals $(0, t')$ and $(0, t'')$ and a long time interval. We shall choose $t' = t'' = 2(\text{on period of a single source})$. In general t' and t'' need not be equal and their choice is found to be relatively insensitive on the accuracy of the approximation procedure [4]. We have also verified this by varying their relative magnitude. The only effect that we have observed is the change of the relative amplitudes of λ_1 and λ_2 . When $t' \leq t''$, $\lambda_1 > \lambda_2$. Otherwise $\lambda_1 < \lambda_2$. Having chosen the values of t' and t'' , the model parameters of the MMPP can be obtained using the following steps-

(1) Let the mean on duration, mean off duration and the interarrival time of cells from a single source be denoted as α^{-1} , β^{-1} and T respectively. The LST of the interarrival time distribution of the renewal process characterizing the arrivals from a source is denoted as $\tilde{f}(s)$ and is given by-

$$\tilde{f}(s) = [1 - \alpha T + \frac{\alpha T \beta}{(s + \beta)}] e^{-sT} \quad (5.8.1)$$

The first three moments of the interarrival time distribution (IAD), evaluated at time t , are denoted as $\mu_1(t)$, $\mu_2(t)$ and $\mu_3(t)$ respectively. These moments are first computed by inversion of their Laplace transforms given by

$$\tilde{\mu}_1(s) = \frac{\lambda}{s^2} \quad (5.8.2)$$

$$\tilde{\mu}_2(s) = \frac{\lambda}{s^2} \left(\frac{1 + \tilde{f}(s)}{1 - \tilde{f}(s)} \right) \quad (5.8.3)$$

$$\tilde{\mu}_3(s) = \frac{\lambda}{s^2} \left(\frac{1 + 4\tilde{f}(s) + \tilde{f}^2(s)}{[1 - \tilde{f}(s)]^2} \right) \quad (5.8.4)$$

where λ is the mean cell arrival rate from a single source and is given by

$$\lambda = \frac{\beta}{T(\alpha + \beta)} \quad (5.8.5)$$

(2) The variance to mean ratio of the IAD evaluated over the intervals $(0, t')$ and $(0, \infty)$ are denoted as $b_{t'}$ and b_{∞} respectively. They are computed using the equations given by-

$$b_{\infty} = \frac{1 - (1 - \alpha T)^2}{(\alpha T + \beta T)^2} \quad (5.8.6)$$

$$b_{t'} = \frac{\mu_2(t') - (\mu_1(t'))^2}{\mu_1(t')} \quad (5.8.7)$$

(3) Let N denote the number of on/off sources that are superposed. The mean arrival rate of cells from the superposed process is denoted as a and is given by $a = N\lambda$. Let $d = q_1 + q_2$. d is computed as the solution of the equation given by-

$$d = \frac{(b_{\infty} - 1)ad^3}{(b_{\infty} - b_{t'})t'} (1 - e^{-dt'}) \quad (5.8.8)$$

Iterative methods like the Newton-Rapson method (see, for e.g. Kreizig[20]) can be used for this purpose.

(4) The parameters of the 2 phase MMPP are finally computed using the equations-

$$q_1 = \frac{d}{2} \left(1 + \frac{1}{\sqrt{4e + 1}} \right) \quad (5.8.9)$$

$$q_2 = d - q_1 \quad (5.8.10)$$

$$\lambda_2 = \left(\frac{ad}{q_2} - \frac{K}{q_1 - q_2} \right) \left(\frac{q_2}{q_1 + q_2} \right) \quad (5.8.11)$$

$$\lambda_1 = \frac{K}{q_1 - q_2} + \lambda_2 \quad (5.8.12)$$

where

$$e = \frac{(b_{\infty} - 1)ad^3}{2K^2} \quad (5.8.13)$$

$$K = \left\{ [at''(at'' - 1)(3\beta t' + at'' - 2) + N\mu_3(t')] - at''^3 - 3a^2t''(b_{\infty} - 1) \left(t'' - \frac{1 - e^{-dt''}}{d} \right) \right\} d^2 \left\{ 3a(b_{\infty} - 1) \left[t''(1 + e^{-dt''}) - 2 \left(\frac{1 - e^{-dt''}}{d} \right) \right] \right\}^{-1} \quad (5.7.14)$$

It may be noted that, knowing the parameters of MMPP 1 and MMPP 2, the model parameters of MMPP 1 and MMPP 2 can be computed using (3.2.1), (3.2.4) and (3.2.5)

TRUNCATION OF THE INFINITE SUMS FOR THE EVALUATION OF $Q'(\omega)$

Infinite summations appear over the indices k and ι in eqns (3.5.3), (3.6.13) and (3.6.14) and they are approximated by finite summations as follows. As the elements of the matrices $P''(n, D)$ and $G_k^{(1)}$ become negligible for large values of n and k , they are computed only for $n = 0, 1, 2, \dots, n_{\max}''$ and for $k = 1, 2, \dots, k_{\max}$ respectively. As x_{ι}'' becomes negligible for large value of ι only finite terms need to be considered in (3.6.14). Let ι_{\max} be the maximum value of ι that needs to be considered. The values of n_{\max}'' , k_{\max} and ι_{\max} are chosen to be the minimum values of n , k and ι for which the following inequalities are satisfied

$$e^T P''(n, D) e < \eta_1 \quad (5.8.15)$$

$$e^T P''(k-1, D) G_{k-1}^{(k-1)} e < \eta_2 \quad (5.8.16)$$

$$x_{\iota}'' e < \eta_3 \quad (5.8.17)$$

where η_1 , η_2 and η_3 are the threshold values for the truncation. For our computations, η_1 , η_2 and η_3 have each been chosen to be 10^{-13}

TRUNCATION OF THE INFINITE DIMENSIONAL MATRICES $Q''(\omega)$ AND $Q'(\omega)$

The invariant probability vectors of $Q'(\omega)$ and $Q''(\omega)$ give the queue length densities of Q1 and Q2, respectively. In view of the coupling between the two queues, these vectors are obtained iteratively. The infinite dimensional matrices $Q'(\omega)$ and $Q''(\omega)$ are truncated to do this. The size of $Q'(\omega)$ which is adequate to ensure a tail probability of less than 10^{-13} when $p_0'' = 1$ is found. The dimension of $Q'(\omega)$ is chosen to be greater than this. Using this as the starting value of the dimension, the QLD of Q1 is found in the actual case where Q2 is non-empty. If the tail probability of Q1 computed thereby is below the desired value, the truncation is adequate. Otherwise the dimension of $Q'(\omega)$ is increased and the QLD is recomputed. This process is repeated until either the required tail probability is obtained or the maximum buffer size for Q1 is reached. It may be noted that in actual numerical implementation of the procedure, due to storage and computational constraints we have to fix the maximum buffer size for both Q1 and Q2. Similarly, the minimum dimension of $Q''(\omega)$ is also determined.

EVALUATION OF NUMERICAL INTEGRALS

For the evaluation of $U_k'(\omega)$ using (3.6.14), we have to evaluate the integrals I_n given by

$$I_n = \frac{1}{D} \int_{w=0}^D P'(n, w) dw \quad (5.8.18)$$

These integrals do not have closed form solutions and hence they have to be numerically integrated. We use the 9 point Newton-Cotes formula (see for eg Hildebrand[21]) for the numerical integral of a function given by

$$\int_{\xi_0}^{\xi_8} f(\xi) d\xi = \frac{4h}{14175} \left[989(f_0 + f_8) + 5888(f_1 + f_7) - 928(f_2 + f_6) \right]$$

$$+ 10496(f_3 + f_5) - 4540f_4 \Big] - \frac{2368}{467775} h^{11} f^{(n)}(\xi) \quad (5.8.19)$$

where h is the step size and $f^{(n)}(\xi)$ denotes the n^{th} differential of $f(\xi)$. By a change of variable of $\xi = \frac{1}{D} w$, eqn (5.8.18) is brought to the form of eqn (5.8.19) and the upper limit for ξ becomes 1. Next, the interval (0,1) is split into 15 intervals and in each interval the function is evaluated at nine points. By repeated application of (5.8.19) in each sub-interval and combining the like terms we get-

$$\begin{aligned} I_n &= \int_{\xi_0}^{\xi_{120}} P'(n, \xi) d\xi = \frac{4h}{14175} \left[989(S_0 + S_8) + 5888(S_1 + S_7) - 928(S_2 + S_6) \right. \\ &\quad \left. + 10496(S_3 + S_5) - 4540S_4 \right] \end{aligned} \quad (5.8.20)$$

$$S_j = \sum_{i=0}^{14} P'(n, (8i+j)hD) \quad \text{for } j = 0, 1, \dots, 8 \quad (5.8.21)$$

As the interval (0,1) is split into 121 points, the step size $h = \frac{1}{120}$ and hence the contribution from the error terms is less than 10^{-22} .

COMPUTATION OF THE EXPONENTIAL MATRICES e^{Q^*nD}

It may be recalled that for the evaluation of β' and μ'_1 , the exponential matrices e^{Q^*nD} have to be computed for $n = 1, 2, \dots, n_{\max}$. They can be evaluated recursively if e^{Q^*D} is evaluated first. For the MMPP models corresponding to the composite traffic from the on/off sources assumed in this thesis, the elements of the matrix Q^*D is less than 1 as D is of the order of μsec . Hence e^{Q^*D} can be evaluated using Taylor series expansion. For the case where Q^* is a 2×2 matrix, closed form expression for this series has been given in (4.4.10). For the case where some of the elements of Q^*D is greater than 1, computation using Taylor series expansion becomes inaccurate and computation intensive (see, for eg Moler[22], [23]). Several methods have been

suggested for overcoming this problem in [23]. For the sake of completion, we describe one of these methods. In the "scaling and squaring" method for computing the exponential of the matrix Q^*D , the exponential of the matrix Q^*D/m is computed first, where m is a power of 2. The value of m is chosen such that the matrix Q^*D/m has all the elements to be less than 1 and hence the exponential of this matrix can be computed accurately using Taylor series. Knowing the exponential of Q^*D/m , the exponential of Q^*D is obtained by squaring the former matrix ℓ times where $\ell = \log_2 m$.

5.9. SIMULATION OF THE PRIORITY SYSTEM

In a simple queueing system like an M/G/1 system with FCFS discipline, closed form expressions for the system probabilities exist and their evaluation requires very little computational effort. On the other hand, for queueing systems with more complex arrival processes and service disciplines, such computationally tractable closed form expressions may not exist. In such cases one may have to evaluate these probabilities numerically. This requires several approximations like performing integrations numerically, replacing infinite sums by finite sums and so on. The system probabilities computed thereby, should be checked against the results obtained using alternate methods. It becomes essential to check the accuracies of the approximations as well as the expressions from which these computations were carried out. We use discrete event simulations for validating the results obtained through numerical computation.

Discrete event simulation of queueing system has been studied in detail in the literature (see for e.g. Fishman[25], Law[26]). It is widely used and is an increasingly popular method for studying complex systems. Most complex real world systems with stochastic elements cannot be accurately described by

a mathematical model that can be evaluated analytically. In such cases, simulation is the only type of investigation possible. However, simulation models are often expensive and time consuming to develop. Each run of a stochastic simulation model produces only estimates of a model's true characteristics for a particular set of input parameters. Several independent runs of the model will typically be required for each set of input parameters to be studied. Compared to results obtained through an analytical approach, simulations would require more time and computational effort for estimating the system characteristics. Hence, to the extent possible, it would be desirable to use simulations as a validation tool - this is what has been done in this thesis to confirm the results obtained through our analysis.

We present a brief summary of the discrete event simulation approach followed for studying a system. In a discrete event simulation model, the state variables of the system are assumed to be changed at discrete points in time as it evolves. At these points in time "events" are said to occur. There could be several types of events and the event type determines which of the state variables will need to be altered. The time between successive occurrence of a particular event is in general random and its probability distribution is assumed to be known. These distributions are, in general, different for different events. Events keep occurring infinitely and as and when an event occurs, the state variables are updated using either the fixed increment approach or the next-event time-advance approach. In the former method, at fixed intervals of time, the "occurrence time" of the various events are checked and the system updated if any of the events occur in that interval. We shall consider here only the latter approach in detail as this is the one used in our simulations. In this a clock known as the simulation clock keeps track of the time that has elapsed since the beginning of the simulation. At the

beginning of the simulation, this clock is initialized to zero. The time of occurrence of various events in the future are also chosen. This is referred to as the "scheduling" of events. The simulation clock is then advanced to the time of occurrence of the most imminent (first) of these events. The state of the system is updated and the data base on the times of occurrence of future events is also updated. Then the simulation clock is advanced to the time of the (new) most imminent event, the system state and the future event times are again appropriately updated. This process of advancing the simulation clock from one event time to another, is continued until finally some pre-specified stopping condition is satisfied. The time that elapses between the beginning of the simulation and its end is referred to as the run time or the replication time of the simulation.

We consider next the various components and organization of a discrete event simulation model -

System state The collection of state variables required to describe the system at a particular time

Simulation clock A variable giving the current value of the simulated time.

Event list A list containing the next time when each type of event will occur

Initialization routine A subroutine to initialize the simulation model at time zero

Counters Variables used to store the statistical information about the system performance

Timing routine A subroutine that determines the next event from the event list and then advances the simulation clock to the time when that event is to occur

Event routine A subroutine that updates the system state when a particular type of event occurs. Every event has its own event routine

Library routines To generate random numbers according to the required probability distribution

Report generator A routine that computes estimates of the desired measures of performance using the counters and reports these when the simulation ends

Main program The routine that invokes the timing routine to determine the next event and then transfers control to the corresponding event routine to update the system state appropriately. The main program may also check for termination and invoke the report generator when the simulation is over. The logical relationship among these components is shown in Fig 5.1

Next, we consider some details on the simulation models developed in this thesis. These models have been used for estimating the busy period distribution (BPD) of the MMPP/D/1 queue with FCFS discipline and for computing the queue length density and queueing delay at the low and high priority queues.

First, the computation of the BPD is considered. It may be recalled that the procedure developed in Sec 4.3 for the computation of the BPD of an MMPP/D/1 queue with FCFS discipline also enables the computation of the BPD of the higher priority queue. Hence we concentrate only on the validation of the BPD computed for the FCFS discipline. We have developed two different simulation models for computing the BPD, one in which the input is generated using on/off sources and the other in which a single MMPP source generates the traffic. As noted in Sec 5.8, the traffic to both Q1 and Q2 are assumed to originate from on/off sources. Hence for the simulation, the input may be generated using the statistics of the individual on/off sources. In view of the mathematical tractability of an MMPP model compared to the on/off source model, we approximated the composite traffic from the on/off sources by an MMPP source in Sec 5.8. Hence, in the second model the input is assumed to be

generated using the statistics of this MMPP source. We shall compare the relative merits of these two simulation models after we describe them in more detail.

SIMULATION WITH ON/OFF SOURCE MODEL

Let the number of constant bit rate (CBR) on/off sources generating the traffic to Q2 be L . Each source is assumed to have exponentially distributed on and off durations as in Sec 5.8. The cells are assumed to be generated periodically in the on period and no cell is generated in the off period. With this source model, the following data bases are updated from time to time during the simulation run.

- (1) `Busy_period_length_count` [N] The i^{th} element of this array denotes the number of times the busy period (BP) was of length i since the beginning of the simulation run.
- (2) `Source_status` [L] The i^{th} element of this array denotes whether the i^{th} source has more cells to generate in its ongoing on period.
- (3) `Cells_count` [L] The i^{th} element of this array gives the number of additional cells to be generated in the current on period of the i^{th} source.
- (4) `Queue_length` No. of cells in the system (both being queued and served).
- (5) `Busy_period_length` Length of the on going busy period, if any.
- (6) `Busy_period_number` Number of BPs completed since the beginning of the run.
- (7) `Busy_period_on` If this flag is 1, it indicates that a busy period is in progress.
- (8) `Terminal_Busy_Period_count` No. of BPs to be completed before terminating the simulation run.

With this model, the various event types to be considered are given in

Fig 8 2 a

Event Type	Description of the event
1	On period of source 1 starts
2	Off period of source 1 starts
3	A Cell arrives from the i th source
4	A cell departs from the queue

Fig 8 2 a The events types for the evaluation of the BPD
through simulation using on/off source model

The tasks to be performed on the occurrence of these events are as follows

Event 1 On period of source 1 starts The tasks to be performed are

- 1 Determine the duration of the on period by calling the random generator and schedule the beginning of the next off period
- 2 Determine the number of cells to be generated in the present on period
- 3 Schedule the next arrival from this source Increment the queue length
- 4 If the queue length is zero, schedule the next departure and make the Busy_period_on flag 1

Event 2 Off period of source 1 starts The tasks to be performed are

- 1 Determine the length of the off period using the random generator and schedule the next on period

Event 3 Cell arrivals for the i^{th} source The tasks to be performed are

- 1 For the i^{th} source if the number of cells remaining to be generated in the on period is greater than 0, schedule the next arrival from this source and update the number of remaining cells to be generated in its on period Increment the queue length
- 2 If the queue length =1, schedule the departure and make the Busy_Period_starts flag 1

Event 4 Departures from the queue The tasks to be performed are

- 1 Decrement the queue length by 1
- 2 If the queue_length equal to 0, make the Busy_period_starts flag 0 Let the busy period length be n, then increment the n^{th} element of the busy_length-count array by 1
- 3 If the queue_length is greater than 0, increment the Busy_Period_length by 1

It may be noted that only one routine is required for each event type for all the sources as these are assumed to be identical

The initialization routine performs the following tasks

- 1 Read the seed for the random number generator, number of busy periods to be completed, source characteristics and the number of sources
- 2 Initialize the various data bases referred to above
- 3 Schedule the on period for all the sources

The main routine performs the following tasks

- 1 Invoke the initialization routine
- 2 Invoke the timing routine
- 3 Invoke the appropriate event routine
- 4 Check for the termination condition If condition not met go to step 2
- 5 Compute the BPD using the Busy_period_length_count array

The busy period distribution is computed using the equation-

$$P[BP=i] = \frac{\text{Busy_period_length_count}[i]}{\text{Terminal_busy_period_number}} \quad (5.9.1)$$

Several runs may be carried out by changing the seed for the random generator and the BPD computed using (5.9.1) for each run is averaged to find the over-all BPD Alternately a single long run may be used to find the BPD The point estimate of the BPD obtained thereby is compared with the 95% confidence

interval of the estimate to check the accuracy of the estimates

SIMULATION OF BUSY PERIOD DISTRIBUTION WITH THE MMPP MODEL

With this model we have to keep track of the phase of the MMPP at the beginning and the end of the busy period. For the purpose of comparison of the results obtained with this model with that obtained through numerical approach, the conditional BPD is estimated i.e., the probability that the BP is of length n and ends in phase j given that the BP started in phase i , is obtained. The probability of the queue being empty at an arbitrary time with phase i is also found. Using this the unconditional BPD is computed. The data base required in this case are as follows

- (1) `Busy_period_length_count [N][2][2]` The $(i,j,k)^{th}$ element of this array denotes the number of times since the beginning of the simulation run that the busy period (BP) was of length i with the phase of the MMPP at the beginning and the end of the busy period as j and k , respectively
- (2) `Phase_of_the_MMPP` the phase of the MMPP
- (3) `Phase_at_the_beginning_of_the_BP` Phase of the MMPP at the beginning of the Busy period
- (4) `Queue_length` No. of cells in the system (both being queued and served)
- (5) `Busy_period_length` Length of the on going busy period if any
- (6) `Busy_period_number` Number of BPs completed since the beginning of the run
- (7) `Busy_period_on` If this flag is 1, it indicates that a busy period is in progress
- (8) `Terminal_Busy_Period__count` No. of BPs to be completed before terminating the simulation run
- (9) `Total_idle_period [1]` The 1^{th} element of this array gives the total

duration for which the queue was empty and the MMPP phase was 1

(10) On_going_idle_period The duration of the idle period in the present phase

With this model, the various event types to be considered for the computation of the BPD are given in Fig 8 2 b

Event Type	Description of the event
1	Beginning of phase 1 of MMPP
2	Beginning of phase 2 of MMPP
3	A Cell arrives in phase 1 or 2
4	A cell departs from the queue

Fig 8 2 b The events types for the evaluation of the BPD
through simulation using the MMPP source model

The tasks performed on the occurrence of these events are as follows

Event 1 Beginning of phase 1 The tasks to be performed are

- 1 Determine the sojourn time in phase 1 using the random number generator and schedule the beginning of phase 2
- 2 Make the phase of MMPP =1, schedule the next arrival with an inter_arrival_time distribution_1
- 3 If queue empty, update the total idle period for phase 2 and make the idle period to be zero

Event 2 Beginning of phase 2 The tasks to be performed are

- 1 Determine the sojourn time in phase 2 using the random number generator and schedule the beginning of phase 1
2. Set the phase of MMPP = 2, schedule the next arrival with an inter_arrival_time distribution_2

- 3 If queue empty update the total idle period for phase 1 and make the idle period to be zero

Event 3 Arrivals in phase 1 or 2 The tasks to be performed are

- 1 Schedule the next arrival with mean inter_arrival_time_i, where i is the phase of the MMPP when this event occurs
- 2 If queue is empty, schedule the next departure and make the Phase_of_BP_start = i and the Busy_period_on flag 1
- 3 If queue is empty, update the total idle period for phase i and make the idle period to be zero
- 4 Increment the queue_length

Event 4 Departures from the queue The tasks to be performed are

- 1 Decrement the queue length by 1
- 2 If the queue_length equal to 0, make the Busy_period_starts flag 0 Let the busy period length be n and the phase of the MMPP at the start of BP and at the present event be i,j , then increment the (n,i,j)th element of the busy_length_count array by 1 and start the counter for the idle period for phase j
- 3 If the queue_length is greater than 0, then increment the Busy_Period_length by 1

The initialization routine performs the following tasks

- 1 Read the seed for the random number generator, number of busy periods to be completed, model parameters of the MMPP
- 2 Initialize the various data bases referred to above
- 3 Schedule the beginning of either phase 1 or phase 2

The main routine performs the following tasks

- 1 Invoke the initialization routine
- 2 Invoke the timing routine

- 3 Invoke the appropriate event routine
- 4 Check for the termination condition If condition not met, go to step 2
- 5 Compute the conditional BPD using the Busy_period_length_count array
- 6 Compute the probability of the queue being empty and in phase 1,2

The conditional busy period distribution is computed using the equation-

$$G_{n_{1j}} = P[BP = nD \text{ sec}, \text{ phase at the end of BP}=j | \text{phase at start}=i]$$

$$= \frac{\text{Busy_period_length_count}[n][i][j]}{\text{No of BPs starting in phase } i} \quad (5.9.2)$$

$$\text{No of BPs starting in phase } i = \sum_{n=1}^N \sum_{j=1}^2 \text{Busy_period_length_count}[n][i][j]$$

$$y_{0i} = P[\text{queue empty and MMPP in phase } i \text{ at an arbitrary time instant}]$$

$$= \frac{\text{Total_idle_period_in_phase } i}{\text{Simulation_clock at the end}} \quad (5.9.3)$$

$$P[BP = nD \text{ sec}] = \frac{y_0 G_n e}{y_0 e} \quad (5.9.4)$$

where $e = [1 \ 1]^T$

Several runs are carried out by changing the seed for the random generator and the BPD computed using (5.9.4) for each run is averaged to find the overall BPD. Alternately, for the probabilities computed using a single run, confidence interval can be computed using the results of Quesenberry[27]. We shall consider more details of this approach towards the end of this section.

Next, we compare the above two models for the evaluation of the BPD. The simulations using the on/off source model require only the source characteristics to be specified. The other model requires the computation of the MMPP parameters using a procedure like Heffes[4] before the simulation can be run. This procedure requires the computation of the inverse Laplace

Transform as discussed in Sec 5.8. Hence the implementation effort required for the computation of the BPD is more with the latter model. However, for the on/off source model, for every source the events 1 and 2 have to be scheduled from time to time. For the MMPP model, these events have to be scheduled for only one source. Hence, for the same traffic, in the on/off source model the event lists need to be processed more number of times. We have implemented both the models to verify the following:

1. The accuracy with which the MMPP is able to approximate the composite traffic from on/off sources.
2. The accuracy of the model parameters obtained through the implementation of the approximation procedure of Heffes [4].

MODEL FOR THE COMPUTATION OF QLDs AND QUEUEING DELAYS AT Q1 AND Q2

In view of the lesser time complexity of the MMPP model, we use only this model for the computation of the QLDs and average queueing delays at Q1 and Q2. As in Sec 5.6, evaluation of the QLD of Q1 at the departure instant of cells from Q1 and those of Q2 at its corresponding departure instants are considered here. As in the last two models, development of the simulation model requires the identification of suitable events and the tasks to be performed on the occurrence of these events. The following data base is required for the simulation model:

- (1) *Queue_length_count_for_Q1* [M][I] The $(m,i)^{th}$ element of this array denotes the number of times the queue length at Q1 system was found to be m and MMPP_1 in phase i at the departure instant of a cell from Q1.
- (2) *Queue_length1* Present queue_length at Q1
- (3) *Queue_length_count_for_Q2* [N][I] The $(n,i)^{th}$ element of this array denotes the number of times the queue length at Q2 system was found to be n

and MMPP 2 in phase 1 at the departure instant of a cell from Q2

(4) *Queue_length2* Present queue_length at Q2

(5) *Busy_period_number* Number of BPs of Q2 completed since the beginning of the run

(6) *Busy_period_on* If this flag is 1, it indicates that a busy period of Q2 is in progress

(7) *Terminal_Busy_Period__count* No of BPs of Q2 to be completed before terminating the simulation

(8) *Total_idle_period_of_Q1[i]* The i^{th} element of this array gives the total duration for which the Q1 was empty and the MMPP_1 phase was 1

(9) *On_going_idle_period_of_Q1* The duration of the idle period of Q1 in the present phase

(10) *Total_idle_period_of_Q2[i]* The i^{th} element of this array gives the total duration for which the Q2 was empty and the MMPP 2 phase was 1

(11) *On_going_idle_period_of_Q2* The duration of the idle period of Q1 in the present phase

(12) *Total_number_of_arrivals_at_Q1* Gives the total number of arrivals at Q1 (since the beginning of the simulation) which are accepted into the queue for service

(13) *Total_number_of_arrivals_at_Q2* Gives the total number of arrivals at Q2 (since the beginning of the simulation) which are accepted into the queue for service

(14) *Buffer_size_of_Q1, Buffer_size_of_Q2* Denote the maximum capacity (in number of cells) of the buffers of Q1 and Q2

(15) *Simulation_run_time_limit* The maximum time for which the simulation is to be run if no other termination condition is met

With this model, the various event types to be considered for the evalua-

tion of the queue length density and queueing delay are given in Fig 8 3

Event Type	Description of the event
1	Beginning of phase 1 of MMPP_2
2	Beginning of phase 2 of MMPP_2
3	Beginning of phase 1 of MMPP_1
4	Beginning of phase 2 of MMPP_1
5	A cell Arrives at Q2
6	A cell arrives at Q1
7	A cell departs from Q2
8	A cell departs from Q1

Fig 8 3 The event types for the evaluation of the QLDs
through simulation using on/off source model

The tasks to be performed on the occurrence of these events as follows

Event 1 Beginning of phase 1 of MMPP_2 The tasks to be performed are

- 1 Determine the sojourn time in phase 1 using the random number generator and schedule the beginning of phase 2 of MMPP_2
- 2 Make the phase of MMPP_2 = 1, schedule the next arrival with an inter_arrival_time distribution_1 of MMPP_2
- 3 If Q2 is empty, update its total idle period in phase 1 and begin its idle period in phase 2

Event 2 Beginning of phase 2 of MMPP_2 The tasks to be performed are

- 1 Determine the sojourn time in phase 2 using the random number generator and schedule the beginning of phase 1
- 2 Make the phase of MMPP_2 = 2, schedule the next arrival with an inter_arrival_time distribution_2 of MMPP_2
- 3 If Q2 is empty, update its total idle period in phase 1 and begin its idle period in phase 2

Event 3 Beginning of phase 1 of MMPP₁ The tasks to be performed are

- 1 Determine the sojourn time in phase 1 using the random number generator and schedule the beginning of phase 2
- 2 Make the phase of MMPP₁ =1, schedule the next arrival with an inter_arrival_time distribution₁ of MMPP₁
- 3 If Q1 empty, update its total idle period in phase 1 and begin its idle period in phase 2

Event 4 Beginning of phase 2 of MMPP₁ The tasks to be performed are

- 1 Determine the sojourn time in phase 2 using the random number generator and schedule the beginning of phase 1
- 2 Make the phase of MMPP₁ = 2, schedule the next arrival with an inter_arrival_time distribution₂ of MMPP₁
- 3 If Q1 is empty, update its total idle period in phase 1 and begin its idle period in phase 2

Event 5 Arrivals at Q2 in either phase 1 or 2 of MMPP₂ The tasks to be performed are

- 1 Increment the queue length of Q2 if it is below the Q2_buffer_limit
- 2 If the queuelength is incremented, login the arrival time
- 3 If Q1 is empty and Q2 has become non-empty just now, schedule the next departure from Q2 and set the Busy_period_on flag as 1
4. If Q2 has become non-empty just now, terminate the idle period of Q2 and update its total idle period in phase 1, where 1 is the phase of MMPP₂ when this event occurs

Event 6 Arrivals at Q1 in either phase 1 or 2 of MMPP₁ The tasks to be performed are

- 1 Increment the queue length of Q1 if it is below the Q1_buffer_limit
- 2 If the queuelength is incremented, login the arrival time

- 3 If Q2 is empty and Q1 has become non-empty just now, schedule the next departure from Q1
- 4 If Q1 has become non-empty just now, terminate the idle period of Q1 and update its total idle period in phase i , where i is the phase of MMPP₂ when this event occurs

Event 7 Departure from Q2 The occurrence of this event is checked by testing the Busy_period_on flag If this flag is 1, a departure from Q2 occurs when the server completes the on-going service and event 7 occurs The tasks to be performed are

- 1 Decrement the queue_length of Q2, If queue_length2=0, then do the following tasks
 - (a) reset the Busy_period_on flag to 0
 - (b) Increment the Busy_period_number
 - (c) Start the idle period of Q2 in phase i , where i is the phase of the MMPP₂ when this event occurs
 - (d) If queue_length1 greater than 0, schedule the next departure from Q1
- 2 Compute the queueing delay for the cell just departed using its arrival time Update the total queueing delay
- 3 Let the queue_length2 be n and the phase of MMPP₂ be i Then increment the $(n,i)^{th}$ element of the queue_length_count_for_Q2

Event 8 Departure from Q1 The occurrence of this event is checked by testing the Busy_period_on flag If this flag is 0, a departure from Q1 occurs when the server completes the on-going service and event 8 occurs The tasks to be performed are

- 1 Decrement the queue_length of Q1, If queue_length1=0, then do the following tasks
 - (a) Start the idle period of Q1 in phase i , where i is the phase of the MMPP₁ when this event occurs

(b) If `queue_length2` greater than 0, schedule the next departure from Q2
and set the `Busy_period_on` flag to be 1

2 Compute the queueing delay for the cell just departed using its arrival
time Update the total queueing delay

3 Let the `queue_length1` be `n` and the phase of MMPP₁ be 1 Then increment the
(`n,1`)th element of the `queue_length_count_for_Q1`

The initialization routine performs the following tasks

1. Read the seed for the random number generator, number of busy periods to be
completed, model parameters of the MMPP₁ and MMPP₂ and buffer_sizes for
Q1 and Q2, `simulation_run_time_limit`

2 Initialize the various data bases referred to above

3 Schedule the beginning of either phase 1 or 2 for both the MMPPs.

The main routine performs the following tasks

1 Invoke the initialization routine

2 Invoke the timing routine

3 Invoke the appropriate event routine

4 Check for the termination condition If termination condition not met, go
to 2

5 Compute the QLDs using the `queue_length_count` arrays

6 Compute the average queueing delays using the `total_queueing_delay` and the
total arrivals statistics

7. Compute the probabilities of Q1 and Q2 being empty in phase 1 and 2

The QLDs of Q1 and Q2 are computed using the equations-

$$\begin{aligned}
 x'_{1j} &= P[\text{Queue length of Q1}=1, \text{MMPP1 in phase } j \text{ at a departure from Q1}] \\
 &= \frac{\text{queue_length_count_for_Q1}[1][j]}{\text{Total_number_of_arrivals_at_Q1}} \quad (5.9.5)
 \end{aligned}$$

$$x''_{1j} = P[\text{Queue length of Q2}=1, \text{MMPP2 in phase } j \text{ at a departure from Q2}]$$

$$= \frac{\text{queue_length_count_for_Q2}[i][j]}{\text{Total_number_of_arrivals_at_Q2}} \quad (5.9.7)$$

The average queueing delays at Q1 and Q2 are computed using the equations-

$$\text{Av_delay_at_Q1} = \frac{\text{Total_queueing_delay_at_Q1}}{\text{Total_nnumber_of_arrivals_at_Q1}} \quad (5.9.8)$$

$$\text{Av_delay_at_Q2} = \frac{\text{Total_queueing_delay_at_Q2}}{\text{Total_number_of_arrivals_at_Q2}} \quad (5.9.9)$$

The probabilities of Q1, Q2 being empty at an arbitrary epoch are given by-

$$\begin{aligned} y'_{01} &= P[\text{Q1 empty and MMPP}_1 \text{ in phase 1 at an arbitrary time instant}] \\ &= \frac{\text{Total_idle_period_at_Q1 in phase 1}}{\text{Simulation_clock at the end}} \end{aligned} \quad (5.9.10)$$

$$\begin{aligned} y''_{01} &= P[\text{Q2 empty and MMPP}_2 \text{ in phase 1 at an arbitrary time instant}] \\ &= \frac{\text{Total_idle_period_at_Q2 in phase 1}}{\text{Simulation_clock at the end}} \end{aligned} \quad (5.9.11)$$

Several runs may be carried out by changing the seed for the random generator and the statistics computed in each run are averaged to find the overall statistics. Alternatively, the simulation may be run for a very long duration with a single starting seed and the above statistics can be computed using a single run, this procedure will give accurate results in the case in which the period of the random number generator is very large, i.e. the number of calls of this routine, after which the starting seed is encountered again, is very large. This can be achieved by using a very large starting seed. In this case, for the probabilities computed using a single run, confidence interval can be computed using the results of [27]. Some details of this approach are given next.

COMPUTATION OF THE CONFIDENCE INTERVALS FOR THE PROBABILITIES

We have considered the computation of the busy period distribution and queue length density using simulation, by computing the appropriate ensemble averages over a short period of time. The length of the busy period is a random variable and out of a total of N busy periods studied, in a simulation, the event that the BP length = n (for $n = 1, 2, \dots$) in a particular BP can be considered as the outcome of a multinomial trial. Similarly, the event that the queue length = n at a departure instant, out of a total of N departures occurring during the run time of a simulation, can also be considered as the outcome of a multinomial trial. Let n_i denote the frequency of occurrence of the i^{th} event out of a total of N trials. Let π_i , for $i = 1, 2, \dots, k$, denote the probability of occurrence of all the possible k events. It is shown in Conover[28] that if $E(n_i) = N\pi_i$ is sufficiently large (typically greater than 5), the statistic

$$\chi^2 = \sum_{i=1}^k (n_i - N\pi_i)^2 \frac{1}{N\pi_i} \quad (5.9.12)$$

is distributed approximately as a chi-square variate with $(k-1)$ degrees of freedom. In Quesenberry[27], it is shown that the confidence intervals for π_i is given by

$$\pi_i = \left[\chi^2 + 2n_i \pm \sqrt{\chi^2 \left[\chi^2 + 4n_i(N-n_i)\frac{1}{N} \right]} \right] \frac{1}{2(N + \chi^2)} \quad (5.9.13)$$

where $\chi^2 = \chi_{\alpha, k-1}^2$ is the upper α percentage of the chi-square distribution and is available in a tabular form for various values of α and k in the literature (see, for e.g. [26], [28]). The confidence interval for BPD and QLDs can be obtained using (5.9.13) for any confidence percentage required for a given maximum queue length or busy period length.

Finally, we consider some implementation issues for all the three models considered above. The first issue concerns the choice of the programming

language for their implementation. In view of the initial training required for using the professional simulation languages and the fact that these were not readily available, we decided against using these for our simulations. Considering the fact that the simulations required by us could be easily set up using general purpose programming languages, we decided to adopt these for implementing our simulator. Here, the choice was between the programming languages FORTRAN and C and we finally settled on using C for our implementation. This choice has the following advantages (see, for e.g. Kernighan[29])

- (1) In view of the better list processing capability of C compared to FORTRAN, the simulation run times are expected to be small with C compared to the latter.
- (2) Even with C, CPU times required for running the simulation model for the QLDs are typically 12-18 hours, on a Convex 220 machine. Since a large range of input traffics are to be studied, managing the CPU time from this machine was often found to be difficult. Hence, running the simulations on some other machines also was felt desirable. Here again, FORTRAN compilers were not available for all the machines accessible to us whereas compilers for C were available on all machines.
- (3) The C language has the advantages of calling the system routine inside its program. For example, the Fortran subroutines can be called and the system commands like "date" can be invoked. This latter command is quite useful for book keeping purposes. For example, to note when a particular simulation started and ended and so on. As a large volume of simulations are run, such labelling is quite useful.
- (4) The "CLOCK" command of C is another attractive feature. It gives the CPU time that has elapsed since the beginning of the program. This command can be

used to find the time required to run the simulation. This is advantageous because we can terminate the simulation run either based on the total CPU time consumed or based on an alternative condition like the number of busy periods completed, whichever happens first. In the absence of this feature, with no *a priori* knowledge about the time required for the run, one has to blindly submit the jobs. In the Convex machine, the CPU run time limits for the quick, short, long and very long queues are typically 30, 120, 480 and greater than 480 minutes respectively. The quick jobs have the highest priority and the very long jobs the lowest priority for receiving service. Submitting a job blindly to a queue, may result in the job getting terminated abortively for exceeding the CPU time limit for that particular queue. Alternately, it may take an unduly long time to run the simulation because it was submitted to a queue for which the maximum time limit was much larger than the actual requirements.

5.10. NUMERICAL RESULTS

The results obtained through numerical computation using the exact model and the approximate model for some typical combination of traffic at the low and high priority queues are presented here. The validation of these results has also been carried out by comparing them with the results obtained through simulation. The programs for the numerical computation have been written in FORTRAN and the simulation routines have been implemented in "C" language. Most of the programs for the numerical computations and the simulations have been executed in a CONVEX -220 machine, other machines such as SUN and HP workstations have also been used.

Some of the assumptions and approximations made for the numerical computations have already been discussed in Sec 5.8. For the numerical

examples considered here, the characteristics of the various on/off sources used are given in Table 5.1

Source type	Average on duration	percentage on duration	on bit rate
Type 1	33 msec	35	1 Mbps
Type 2	3.3 msec	35	1 Mbps
Type 3	1 msec	35	1 Mbps
Type 4	50 μ sec	10	10 Mbps
Type 5	5 μ sec	1	100 Mbps
Type 6	15 μ sec	10	100 Mbps

Table 5.1 Characteristics of the different constant bit rate on/off sources

The on/off source whose parameters are given in the i^{th} row of this table is denoted as the Type i source. For ease of reference, we shall denote the exact model and the approximate model as model I and model II respectively. Let N_1 , N_2 denote the number of on/off sources generating the traffic to Q_1 and Q_2 , respectively. The steps involved in the computation of the QLDs are as follows:

(1) The types of on/off sources generating the traffic to Q_1 and Q_2 and the numbers N_1 and N_2 are fed as the inputs to the routine. Knowing these information, the model parameters of both MMPP 1 and 2 are obtained using the Heffes approximation procedure [4]. (The detailed set of equations required for this purpose have been given in Sec. 5.8.)

(2) Using the recursive procedure given in Sec. 4.4, the matrices $P''(n, D)$ and $P'(n, D)$ for $n = 0, 1, 2, \dots$ are found. The truncation indices n''_{\max} , n'_{\max} are determined using (5.8.15). The row sums of the $P''(n, D)$ for $n = 0, 1,$

n_{\max}'' are found and it is verified that they sum up to 1. The corresponding sums for $P'(n,D)$ are computed and their sum is also verified to be 1.

(3) G_k'' s and $G_k^{(m)}$ are computed using the recursive procedure given in Sec 4.3.

The value of k_{\max} is computed using (5.8.16). The row sums of G_k'' for $k = 1, 2, \dots, k_{\max}$

are formed and verified to be 1. The matrices $P'(n,kD)$ for $k = 1, 2, \dots, k_{\max}$

are found using the recursive procedure given in Sec 4.4. The maximum value of n required for each value of k is determined using (5.8.15) with D replaced by kD . The row sums of $P'(n,kD)$ are found and verified to be 1.

(4) The integrals $\frac{1}{D} \int_{w=0}^D P'(n,w)dw$ for $n = 0, 1, \dots, n_{\max}$ are computed numerically using (5.8.20).

(5) Corresponding to the case where $p_0'' = 1$, the value of y_0' and p_0' are obtained by solving the resulting single priority queue. The initial value of y_0' for the prioritized queue is set to be this value. The initial value of y_0'' is chosen to be the null vector.

(6) The matrices $A_m''(\omega)$, $U_k''(\omega)$, G'' , $e^{\underline{Q}^* n D}$ and the vectors β'' , π are precomputed and stored.

(7) Knowing the value of y_0' , the matrices $B_n''(\omega)$ are computed for $n = 0, 1, \dots, n_{\max}$.

If the iteration number is 1, the size of $Q''(\omega)$ is chosen to be greater than that required with the FCFS discipline to ensure a given tail probability. Otherwise, the size determined in the previous iteration for Q2 is used as the starting value. The QLD of Q2 is computed. If the smallest probability computed is greater than the required tail probability, the dimension of $Q''(\omega)$ is increased and the QLD is recomputed. This procedure is repeated until either the required tail probability is obtained or the maximum buffer size limit for Q2 is reached.

(8) The vector y_0'' is computed next. If it is to be computed by computing x_0'' using (5.2.2), then μ_1'' and k_0'' are computed using the equations given in Sec 4.5 and 4.6. Otherwise the value of x_0'' , computed through the iterative procedure is used. μ_2'' is computed using (4.6.4) and finally y_0'' and p_0'' are computed using (5.5.9) and (5.5.10).

(9) The value of ι_{\max} is determined using (5.8.17). The matrices $C(k)$ for $k = 1, 2, \dots, k_{\max}$ and $F(\iota)$ for $\iota = 0, 1, \dots, \iota_{\max}$ are computed. The matrix G' is computed using the equations given in Sec 4.2.

(10) The matrices $A'_m(\omega)$, $U'_k(\omega)$ and $B'_m(\omega)$ are computed next. If the iteration number is 1, the size of $Q'(\omega)$ is chosen to be greater than that required with the FCFS discipline to ensure a given tail probability. Otherwise, the size determined in the previous iteration for Q1 is used as the starting value. The QLD of Q1 is computed. If the smallest probability computed is greater than the required tail probability, the dimension of $Q'(\omega)$ is increased and the QLD is recomputed. This procedure is repeated until either the required tail probability is obtained or the maximum buffer size limit for Q1 is reached.

(11) The vector y_0' is computed next. If it is to be computed by computing x_0' using (5.2.3), then μ_1' and k_0' are computed using the equations given in Sec 4.5 and 4.6. Otherwise the value of x_0' , computed through the iterative procedure is used. μ_2' is computed using (4.6.5) and finally, y_0' and p_0' are computed using (5.5.30) and (5.5.31).

(12) The largest difference between y_0'' computed in the present iteration and that computed in the previous iteration is found. The corresponding difference is found also for y_0' . If either of these differences is greater than the desired accuracy, steps 7 - 12 are carried out again. Otherwise, the procedure is terminated.

It may be noted that the above procedure is also valid for the approximate model and in this case, the appropriate equations from Sec 5.7 are used.

Since this computational procedure involves several intermediate steps, testing the correctness of each step becomes essential. Checking the row sums of the various stochastic matrices is one step that is carried out for this purpose. Since the computation of $P(n,D)$ s and G_k 's are the crucial steps for the evaluation of the QLDs, we next consider the validation of the computational procedure used for the evaluation of G_k 's or equivalently the busy period distribution of Q2. The results obtained through the numerical approach are compared with the simulation results as follows.

As mentioned in Sec 4.3, given that the busy period (BP) starts with a single cell, the BP of Q2 is the same as that of the MMPP/D/1 queue with FCFS discipline with the same arrival statistics. Hence, we concentrate on the busy period of the latter queue. We shall denote this queue as Q.

We assume the traffic to Q to be generated by N_s type 1 on/off sources. The parameters of the MMPP, the matrices $P(n,D)$ and G_k are found as in steps (1) -(3) given above. The probability mass function (PMF) of the busy period length, i.e., the probability that busy period is equal to nD sec, for $n = 1, 2, \dots$ are evaluated using (4.3.13). The PMF obtained corresponding to $N_s = 45, 90$ and 150 are plotted in Fig 5.4, Fig 5.5 and Fig 5.6 respectively. In these figures the busy period length is expressed in units of the cell transmission time (D). The average traffic offered at the FCFS queue (ρ) is computed as $\pi\lambda D$, Here π is the stationary vector of the rate process of the MMPP and λ is the vector denoting the diagonal elements of the arrival rate matrix, Λ of the MMPP.

For obtaining the PMFs through simulation, both the on/off source model

and the MMPP source model, discussed in Sec 5.8, are used. We shall denote these models as Simulation Models A and B respectively. For the computation of the busy period distribution (BPD) of the on/off source model (Simulation Model I), only the no. of sources and the source characteristics need to be specified. For the MMPP model (Simulation Model B), the MMPP model parameters computed through the numerical approach is fed as the input. If the BPD obtained through both the simulation models are to agree, two conditions have to be satisfied. These are (i) the MMPP model parameters have been obtained correctly and (ii) the approximation procedure used to obtain the parameters is accurate for the given on/off source. As mentioned in Sec 5.8, one reason for using the two different simulation models is to test these two conditions. Another reason is to compare their relative time complexity.

The BPDs are obtained using both model A and B for $N_s = 45, 90$, and 150 and the results are plotted in Fig 5.4, Fig 5.5 and Fig 5.6 respectively. From these figures, it can be concluded that the BPD obtained through computation and the Simulation Models A and B agree well. The point estimates of the probabilities are obtained by considering total busy periods of the order 10^8 . The replication time of the simulation routine is chosen accordingly. The 95% confidence interval for the probabilities computed using simulation is found to be within 1-15 % of the point estimates in the entire range. However, for the sake of clarity, we have not shown them in these figures. The run time required for Simulation Model A has been found to be about 1.5 times more than that required for Model B. In view of this, for the computation of the QLDs, only the latter model has been chosen. We have checked the validity of the BPD only upto probabilities of the order of 10^{-5} using simulation. However, we have actually computed the BPD upto tail probabilities of the order of 10^{-15} . The BPDs obtained corresponding to the traffic offered (ρ) at Q2 for

less than 0.5 and greater than 0.5 are shown in Fig 5.7 and Fig 5.8 respectively. From these figures, it can be concluded that the storage required for computing the BPD and QLDs become extremely high as the traffic becomes large.

Having thus verified the accuracy of the procedure for the computation of the BPD and hence for the computation of $P(n,D)$, the remaining steps for the computation of the QLDs are checked next. For some typical examples, the QLDs of both Q_1 and Q_2 are computed using both Gaussian elimination and Toeplitz matrix inversion methods discussed in Sec. 5.6. The results obtained using both the methods have been found to agree well. In view of the lesser computational requirements of the latter method, for all the subsequent computations only that method has been chosen. As an additional computational check, the QLDs are obtained using both the methods for computing y_0'' and y_0' given in steps 8 and 11. Both the methods have been found to give the same result. As the evaluation of x_0'' and x_0' using (5.12) and (5.13) require more computational effort compared to the alternate approach mentioned in step 8 and 11, only the latter approach is chosen for the subsequent computations. Next, the initial vector y_0' used for the iterative procedure is varied and its effect on the convergence of the recursive procedure is studied. The iterative procedure has been found to be relatively insensitive to the choice of the initial vector and is found to converge fast. For a relative accuracy of 10^{-13} , about 6 to 8 iterations have been found to be adequate for all the initial vectors chosen. The list of computational checks that have been discussed so far verify only the *correctness* of the *implementation* of the computational procedure. The *accuracy* of the computational procedure is next checked by comparing these results with that obtained using simulation for several examples.

First we consider the case where the Type 1 sources generate the traffic

to both Q1 and Q2. Let N_2 , the number of sources generating the traffic to Q2, be chosen to be 45. The average traffic offered (ρ) at Q2 becomes 0.1. Next N_1 , the number of sources generating the traffic to Q1 is varied. The QLDs of Q1 and Q2 computed using both the exact model (Model I) and the approximate model (Model II) have been found to be in agreement over a wide range of traffic. Of these examples, for a particular case where $N_1 = 300$ and $N_2 = 45$, the QLDs of Q1 and Q2 obtained are shown in Fig 5.9 and Fig 5.10. (In this case the average traffic offered at Q1 and Q2 become 0.7 and 0.1 respectively). Since both the computational models give the same result only one curve is shown corresponding to computation. The results obtained through simulation are also shown in these figures. From these figures, it can be concluded that the simulation results agree well with those of computation for this example.

Next, we examine the effect of the characteristics of the on/off sources on the accuracy of the results obtained. Fixing the average traffic at Q1, Q2 to be 0.7 and 0.1, the values of N_1 and N_2 required to generate this traffic combination are found for the source types 2-5. Both the type 2 and type 3 sources require N_1 and N_2 to be 300 and 45. The QLDs obtained through simulation and the approximate model corresponding to these two sources are shown in Fig 5.11-5.14, here Figs 5.11 and 5.12 correspond to type 2 sources and Figs 5.13 and 5.14 correspond to type 3 sources. For both the type 4 and type 5 sources, N_1 and N_2 have been found to be 105 and 15 for the given values of average traffic 0.7 and 0.1 to Q1 and Q2, respectively. The QLDs corresponding to these examples are shown in Fig 5.15-5.18, here Figs 5.15 and 5.16 correspond to type 4 sources and Figs 5.17 and 5.18 correspond to type 5 sources. Both the exact (I) and approximate (II) methods give identical results, therefore, only one curve has been shown for the computational

results in Figs 5 11-5 18. It can be also be seen from these figures that the computational results agree well with simulations. Only slight deviations between the computed and simulated results are observed at the tails of the distribution for Q1, the lower priority queue. From these figures it can be concluded that (a) the analytical model agrees well with simulations over a wide range and that (b) for the analytical approach, the approximate model is sufficiently accurate for these examples. Similar comments can also be made about the results shown in Figs 5 19 and 5 20 for Type 1 sources corresponding to the case where $N_1 = 150$ ($\rho = .35$) and $N_2 = 90$ ($\rho = 0.21$). In general, the QLDs of Q1 and Q2 computed using the approximate model for low to medium traffic at Q1 been found to be in close agreement with those obtained through the exact model and through simulations.

At first sight, the above examples do give the impression that the approximate model is probably just as good as the exact model. The fact that this is not the case is demonstrated through examples next. Based on these examples and our insight into the functioning of this system, we also empirically provide information on the situations where we can expect that the exact analytical model (I) should be used rather than the simpler approximate model (II).

We re-examine the results of Figs 5 9 - 5 20 for this purpose. In these figures, the traffic offered at Q1 in an interval of one service time (D) when the MMPP 1 is in either phase 1 or 2 as well as the transition rate of the state of the MMPP 1 have also been shown. The corresponding quantities for Q2 in phase 1 and 2 of MMPP 2 are also shown. From the values of the " ρ "s corresponding to MMPP 2 in phases 1 and 2 as given in these figures, it can be seen that the " ρ "s in different phases of a given MMPP are of the same order. Hence, one obvious case that needs to be tested is where the " ρ " in one phase

is significantly higher than that in the other phase

In order to test this, we next consider a case in which the composite traffic at Q2 has the arrival rate in one phase to be ten times higher than that in the other phase. The average traffic offered at Q2 is set be 0.1. The traffic to Q1 is assumed to originate from 300 type 1 sources with average traffic as 0.7. In this case, the buffer size of Q1 and the computational times required for obtaining tail probabilities at Q1 to be as low as 10^{-13} turned out to prohibitively high. We therefore obtained the QLDs by limiting the buffer size at Q1 to be 475. With this, the results obtained using model I (exact), model II (approximate) and the results from simulations are plotted in Figs 5.21 and 5.22. Fig 5.21 shows the QLD for the low priority queue Q1 and Fig 5.22 shows the QLD for the high priority queue Q2. The computational results differ for the QLD of Q1 as shown in Fig 5.21. The QLDs obtained through the two computational approaches still give similar results for Q2 - hence only one computational curve has been shown in Fig 5.22. In this example, the results obtained using the exact model agree well with simulations in all cases. However, the approximate method does not give the right results for the QLD of Q1.

We consider another example in which the traffic arrival rates in the two phases of MMPP₂ are drastically different. For the type 6 sources, the composite traffic from 16 sources has the required burstiness characteristics and the average traffic intensity of 0.1. We again assume that the traffic to Q1 is generated by 300 type 1 sources. Assuming an infinite buffer at Q1, the QLD of Q1 obtained using model I (exact), model II (approximate) and simulations are plotted in Figs 5.23 and 5.24. In this example also, the QLD of Q1 obtained using the exact model and simulation agree well but differ significantly from the results obtained for the QLD of Q1 using the approximate model. From these two examples, we can conclude that the

approximate model may not be accurate enough for situations where the composite traffic at Q2 has a large burstiness. On the other hand, the exact model is able to give accurate results which agree with the simulation results even when the traffic to Q2 is bursty. For these examples, the QLDs of Q2 obtained using computation and simulation are shown in Fig 5.22 and 5.24. These are found to be generally in good agreement. (We explain later why the results of Q2 using the approximate analytical model agree well in these examples even though the results for Q1 do not.)

In order to consider the situations where the queues are heavily loaded, we give examples next where the total traffic offered to the server at Q1 and Q2 is close to its capacity. Most of these examples require very large storage and computational resources for computing the QLDs at Q1 with tail probabilities as low as 10^{-13} . To make our numerical computations easier, we assume the buffer size at Q1 to be finite in most of these examples. The buffer size of Q2 is considered to be infinity in all these examples.

In the first set of examples, assuming the traffic to Q1 and Q2 to originate from (350,45) type 1 sources, the QLDs are obtained using all the three methods. The QLDs of Q1, the lower priority queue, corresponding to buffer sizes of 400 and 100 are shown in Fig 5.25 and 5.26, respectively. The QLDs of Q2 corresponding to both the buffer sizes have been found to be essentially same and the results using all the three methods are shown in Fig 5.27 for the case where the Q1 buffer size is 400. In the next set of examples, the traffic to Q1 and Q2 are assumed to originate from (360,45) type 1 sources. The QLDs are obtained using all the three methods for Q1 buffer sizes of 475 and 150. The QLD of Q2 again has been found to be insensitive to the Q1 buffer size and is plotted in Fig 5.28 for a buffer size of 475 for Q1. The QLDs of Q1 corresponding to both the buffer sizes are shown in Fig 5.29 and 5.30. In

the above two sets of examples, the average traffic offered at Q2 has been chosen to be 0.1

Next, we consider four examples in which the average traffic offered at Q2 is 0.2. In the first example, the traffic to Q1 and Q2 are assumed to originate from (300,90) type 1 sources. The QLDs of Q1 obtained corresponding to buffer sizes of ∞ and 100 are depicted in Fig 5.31 and 5.32. The QLD of Q2 is again found to be insensitive to the buffer size of Q1 and is shown in Fig 5.33 for the case where the Q1 buffer size is ∞ . In the second example, the traffic to Q1 and Q2 is assumed to originate from (320,90) type 1 sources. The QLDs at Q2 and Q1 are computed using all the three methods for a Q1 buffer size of 475 and are shown in Fig 5.34 and Fig 5.35, respectively. In these figures, the QLDs obtained assuming the Q2 traffic to be from a Poisson process with average traffic intensity of 0.2 are also shown. In the third example, the traffic is assumed to originate from (300,90) type 2 sources and in the fourth example the traffic originates from (300,90) type 3 sources. The QLDs obtained corresponding to these two examples, assuming infinite buffer sizes for both Q1 and Q2, are shown in Figs 5.36-5.39.

Finally, we consider an example in which the traffic to Q1 and Q2 originate from (240,135) Type 1 sources. The average traffic offered at Q1 and Q2 become 0.56 and 0.31 respectively. The QLDs at Q1 and Q2 computed using all the three methods are shown in Fig 5.40 and Fig 5.41. In this case we have assumed both Q1 and Q2 to have infinite buffers.

In all the seven set of examples considered above, two conditions are consistently found to be satisfied. Firstly, the QLDs of Q1 computed using simulations agree well with that obtained using the exact model (Model I) but differs significantly from those obtained using the approximate model (Model II), This is especially true at higher queue lengths. Secondly, the QLDs of Q2

obtained using Models I and II and simulation agree well in all these examples. Our reasons for expecting these results are given next.

We have noted in the above examples that the approximate model fails to compute the QLD of Q1 accurately in cases where the Q2 traffic is either highly bursty or the total traffic offered to the server is close to the capacity. We have indicated the traffic offered (ρ) at Q2 in each phase of MMPP 2 and that offered at Q1 in each phase of MMPP 1 in Figures 5.21 - 5.41. In all these figures, the total traffic offered to the server becomes close to and even greater than the capacity of the server ($\rho = 1$) when both the MMPPs happen to be in the maximum arrival phase (phase 1, in the examples considered above). From the phase transition rates indicated in these figures, it can be noted that the probability of occurrence of this phase pair is not negligible. It may be recalled that in the approximate model the busy period distribution (BPD) of Q2 has been obtained as the weighted average of the BPD corresponding to each possible initial phase of MMPP 2. Hence as far as Q1 is concerned, the approximate model treats the mean arrival rate at Q2 to be constant w.r.t. time. Let us assume that the phases of MMPPs are labelled such that phase 1 is the one with the larger arrival rate (compared to phase 2). When both the MMPPs are in phase 1, the server is fed with more number of arrivals than that predicted by Model II. The queue length at Q1 is more than what is predicted by Model II in this case. Similarly, when both the MMPPs are in phase 2, there will be less number of arrivals than what is predicted by Model II. In this case, the queue lengths at Q1 is less than that predicted by Model II. When the traffic offered at Q1 and Q2 are low to medium, the increase in the queue lengths at the former intervals of time are compensated by the decrease in the queue lengths at the later intervals of time. Hence the statistical averages of the queue lengths turn out to be the same for both the exact and the approx-

approximate model. A different scenario emerges when the MMPP 2 is either highly bursty or the server is loaded closed to its capacity. It may be noted here, that the increase in the queue length, with increase in the traffic offered (ρ) to the server, is an increasingly non-linear function of ρ . Hence at loads closed to the capacity of the server, the increase in the queue lengths at Q1 when the MMPPs are in phase 1 become much larger than the decrease in the queue lengths at Q1 when both the MMPPs are in phase 2. Because of this, the actual statistical average of the queue lengths at Q1 as computed by model I become larger than that computed by model II. Hence in this range of traffic, model II under estimates the queue length.

Next we examine why the QLDs of Q2 agree in all the three methods. It may be recalled here that the computation of the QLD of Q2, depends only on p'_0 and not on any other parameter of Q1. In all the nine examples that we considered above, at lower queue lengths the QLDs of Q1 computed using both model I and II, are equal. When the QLDs computed using model I and II, start departing, the probabilities (QLDs) are below 10^{-2} and hence the p'_0 computed using both the models are expected to be essentially the same. We have plotted the variation of p'_0 with variation in the traffic offered at Q1 as a function of the traffic offered at Q2 in Fig 5.42. The four set of plots shown in this figure are obtained by keeping the traffic offered at Q2 to be 0.05, 0.1, 0.2 and 0.3 respectively. From this figure it can be concluded that p'_0 computed using both the models agree. The first consequence of this observation is that the QLDs of Q2 obtained using both model I and II match. Secondly, if one is interested only in the moments of the queue lengths at Q1, then one can compute p'_0 using the approximate model. We have discussed in Sec 5.6 how p''_0 and p'_0 can be evaluated without actually evaluating the QLD of Q1. Hence, using the approximate model, the moments of the queue lengths at Q1 can be found, this

will be faster than finding this with the exact model. Thirdly, we shall find in Chapter 6, that the computation of the average queueing delays at Q1 and Q2 require only y'_0 and y''_0 . Hence the approximate model can also be used for computing the average delays at Q1 and Q2 quickly.

As noted in Sec 5.7, the approximate model can also be used to compute the QLDs when either Q1 or Q2 or both have Poisson arrivals. As noted earlier, the approximate model approximates the busy period distribution of Q2 to be the weighted average of that corresponding to each possible initial phase. Hence as far as Q1 is concerned, the Q2 arrivals look like "Poisson". However, for Q2, this model assumes the traffic to be MMPP. Another approximation would be to approximate the traffic at Q2 to be Poisson and obtain its arrival rate as the weighted average of the arrival rates of MMPP 2. The QLDs of Q1 and Q2 computed using this approximation corresponding to the case when the traffic to Q1 and Q2 originate from (320,90) type 1 sources are shown in Fig 5.43 and Fig 5.44. The QLDs computed using the exact, approximate and the Poisson models corresponding to the case when the traffic originates from (240,135) type 1 sources are shown in these figures. The QLD of Q1 computed using both Model II and Poisson model match perfectly but depart from the results computed using the exact model. The QLD of Q2 computed using both Models I and II agree all the way upto probabilities as low as 10^{-12} . The QLD of Q2 computed using the Poisson model agree well with the results obtained using Models I and II at lower queue lengths but differs from them at the higher queue lengths. Hence, if the computation of the moments of the queue lengths are the basic parameters of interest, the Poisson model may also be used in place of Model II.

At lower queue lengths the accuracy of the results obtained using Models I and II have been checked by carrying out the simulation. The probability of

occurrence of the higher queue lengths are very small and hence they cannot be evaluated using simulation unless prohibitively long run lengths are used. Hence, using an alternate method we examine their validity at higher queue lengths in the next two examples. In the first example the traffic at Q1 and Q2 is assumed to originate from (300,45) type 1 sources. The QLDs obtained using Model I is compared with the case where Q1 and Q2 have dedicated servers in Fig 5.45. As is to be expected, the QLD of Q2 in the prioritized system differ from that corresponding to the case with dedicated server only marginally. The QLD of Q1 departs significantly from that corresponding to the case where it has a dedicated server. In the second example, the traffic originating from (150,90) type 1 sources are considered and the results are plotted in Fig 5.46. The same conclusions are valid for this example as well.

Based on the examples that we have considered so far and a large number of other examples that we have studied, the following conclusions can be drawn. It appears that the approximate model (Model II) should not be used if either of the following two conditions (a) or (b) are true- in these cases, the exact model (Model I) is recommended for the computation of the QLDs.

(a) If $\lambda_1'' / \lambda_2''$ is significantly greater than 1, where λ_1'' , λ_2'' are the arrival rates of the MMPP_2 in the two phases and are labelled such that $\lambda_1'' > \lambda_2''$.

In other words if the traffic to Q2 is highly bursty.

(b) There is a particular phase pair of the two MMPPs which will tend to overload the server and this phase pair is one which is fairly likely to arise.

Otherwise, the simpler computations of the approximate approach (Model II) may be more attractive. At the cost of a marginal degradation in the accuracy of the QLD of Q2, the Poisson model may also be used in place of Model II, resulting in further reduction in the computational effort.

When the moments of the queue lengths are the only parameters of interest, then either Model II or the Poisson model for Q2 may be used for computing the p'_0 and p''_0 required for this purpose. When the moments of the queueing delays are required then Model II may be used for computing the y'_0 and y''_0 required for this purpose.

When the traffic offered to Q2 becomes greater than 0.3, the computational and storage resources required for the computation of the busy period and additional busy period distributions of Q2 become very high. Hence, for such high loads, we may have to assume the buffer size at Q2 to be finite. The details of the computation of the QLDs with a finite sized Q2 is considered in Chapter 7.

REFERENCES.

- 1 M. F. Neuts, "Structured stochastic matrices of the M/G/1 type and their applications", Marcel Dekker, New York, 1989.
- 2 Hunter, J. J., "On the moments of Markov renewal processes", Adv. Appl. Prob. 1, 1969, pp. 188-210.
- 3 V. Ramaswamy, "The N/G/1 queue and its detailed analysis," Adv. Appl. Prob., Vol. 12, pp. 222-261, Mar. 1980.
- 4 H. Heffes and D. M. Lucantoni, "A Markov Modulated characterization of Packetized voice and Data traffic and Related Statistical Multiplexer Performance", IEEE J. SAC, No. 6, pp. 856-867, Sep. 1986.
- 5 R. W. Wolff, "Stochastic modelling and theory of queues", Prentice Hall, New Jersey, 1988.
- 6 J. Medhi, "Stochastic processes", Wiley Eastern Ltd., New Delhi, 1991.

- 7 Udo R Kreiger, Bruno Muller-Clostermann and Michael Sczittnick, " Modeling and analysis of communication systems based on computational methods for Markov chains", IEEE Journal on Selected Areas in Commn , Dec 1990, pp 1630-1648
- 8 William J Stewart, "Numerical solution of Markov chains", Marcel Dekker, Inc , New York, 1991
- 9 V Ramaswami, "Stable recursion for the steady state vector for Markov chains of M/G/1 type", Stochastic models, 4, 1988, pp 183-188
- 10 Lucantonı D M , " New results on the single server with a batch Markovian arrival process", Commun statist - Stochastic models, 7(1), 1991, pp 1-46
- 11 C Blondia , " The N/G/1 finite capacity queue", Commun Statistic Stochastic Model,5(2),pp 373-294 (1989)
- 12 D V Ouellette,"Schur complements and statistics", Linear algebra and its applications 36, March 1981 pp 187-295
- 13 H Akaike, "Block Toeplitz matrix Inversion", SIAM J Appl Math , Vol 24, pp 234-241, March 1973
- 14 Trench W F , "An algorithm for the inversion of finite Toeplitz matrices", Journal of the Society for Industrial and Applied Mathematics, 12, 1964, pp 515-522
- 15 Trench W F , "Inversion of Toeplitz band matrices", Math Computation, Vol 28, 1974, pp 1089 - 1095
- 16 Jain A K , "Fast inversion of banded Toeplitz matrices by circular decompositions", IEEE Trans on Acoustics Speech and Signal Processing, 26, 1978, pp 121-126
- 17 Forsythe G E , Malcolm M A , and Moler C B , "Computer methods for mathematical computations", Prentice-Hall, Inc , 1977

- 18 Barlow and Prostian, "Mathematical theory of reliability" , Wiley, New York, 1965
- 19 B Venkataramani, Sanjay K Bose and K R Srivathsan, "Queue length density and busy period distribution of MMPP/D/1 queue with non-preemptive priority for use in ATM networks", Proc ITC seminar, Bangalore, pp 121-128, Nov-1993, India
- 20 V Ramaswamy, Private communication - Nov 1993
- 21 Kreizig, "Advanced Engineering Mathematics", Wiley Eastern Ltd , New Delhi, 1983
- 22 F B Hildebrand, "Introduction to Numerical Analysis", Second Edition, Tata-McGraw Hill Publishing Company Ltd , New Delhi, 1974
- 23 C Moler and C Van Loan, "Nineteen dubious ways to compute the exponential of a matrix", SIAM Rev 20 (1978) pp 801-836
- 24 C Van Loan, "A note on the evaluation of matrix polynomials", IEEE Trans on Aut Control, Vol AC 24, No 2, 1979, pp 320-321
- 25 G S Fishman, "Principles of Discrete event simulation ", John Wiley, New York, 1978
- 26 A M Law and W D Kelton, "Simulation Modelling and Analysis", Second Edition, McGraw-Hill, Inc , 1991
- 27 C P Quesenberry and H C Hurst, "Large sample simultaneous confidence intervals for multinomial proportions", Technometrics, Vol 6, No 2, May 1964, pp 191-195
- 28 W J Conover, "Practical nonparametric statistics", Second Edition, John Wiley, New York, 1980
- 29 B W Kernighan and D M Ritchie, "The C Programming language", Second Edition, Prentice Hall, Englewood Cliffs, N J , 1988

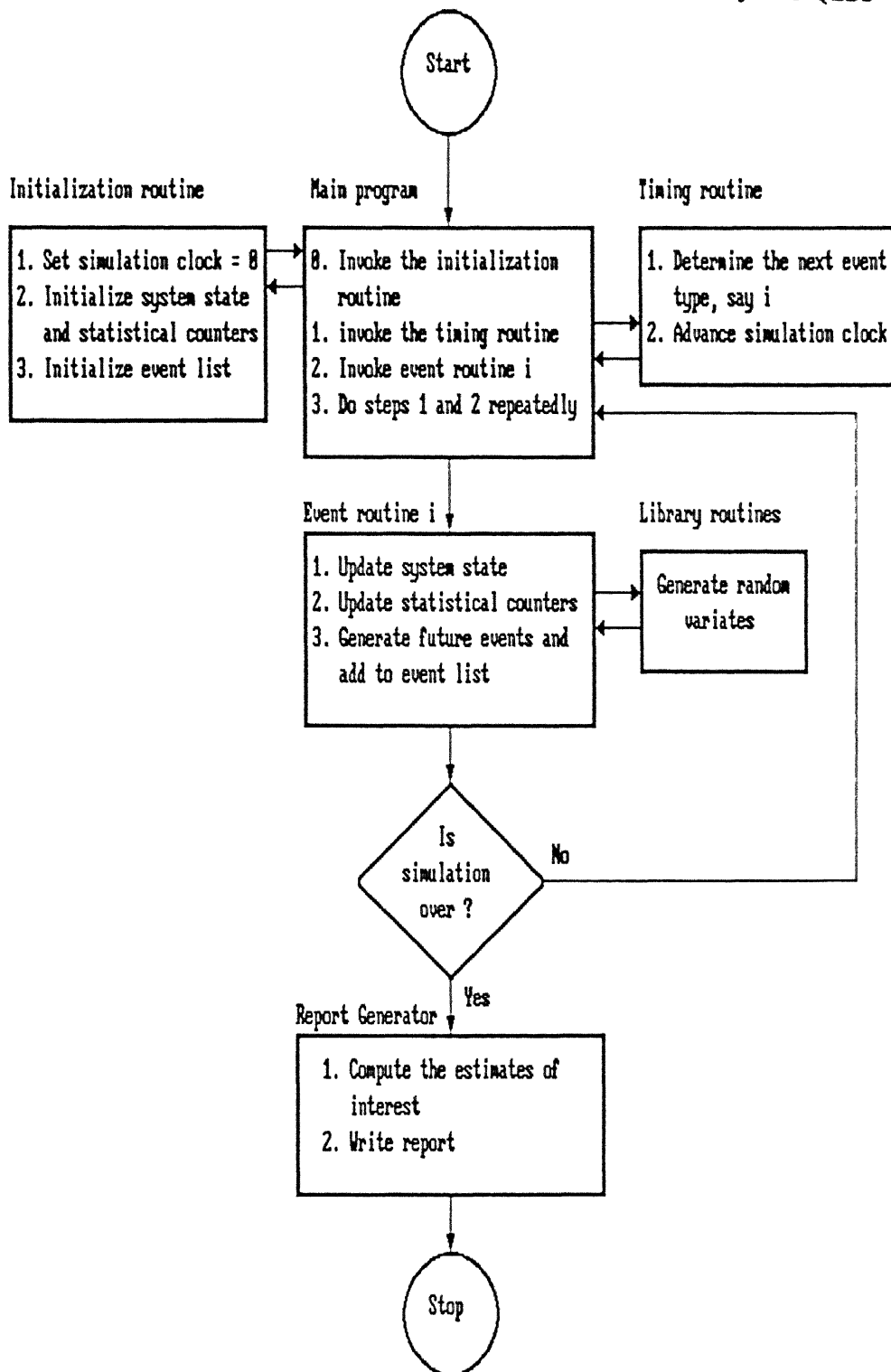


Fig.5.1. Flow of control for the next-event time advance approach

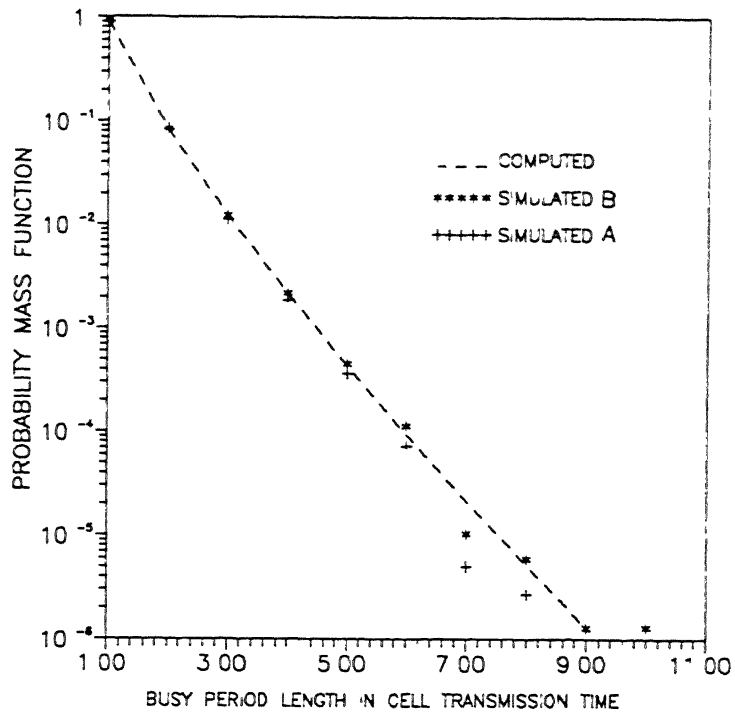


FIG 5.4 P.M.F OF BUSY PERIOD LENGTH OF MMPP/D/1 QUEUE
COMPUTED AND SIMULATED FOR $N_s = 45$ ($\rho = 0.10$)

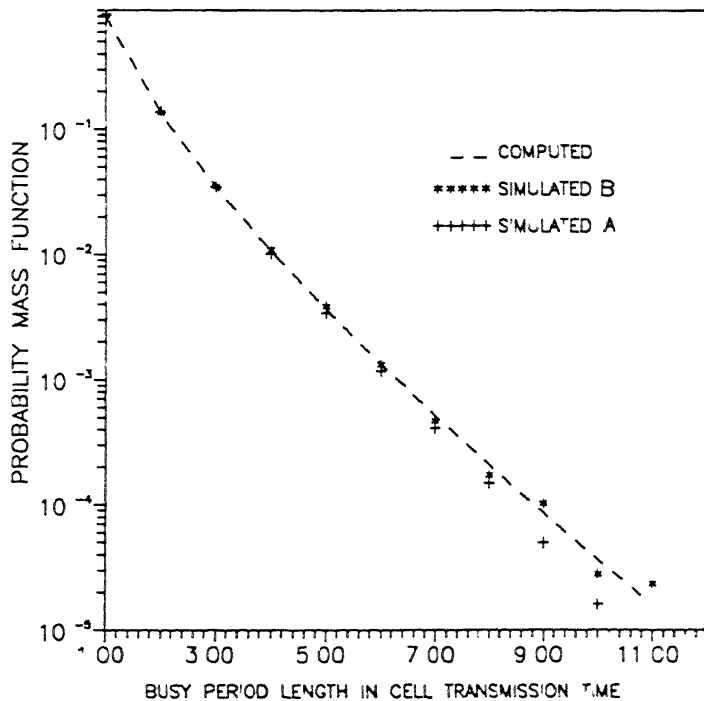


FIG 5.5 P.M.F OF BUSY PERIOD LENGTH OF MMPP/D/1 QUEUE
COMPUTED AND SIMULATED FOR $N_s = 90$ ($\rho = 0.21$)

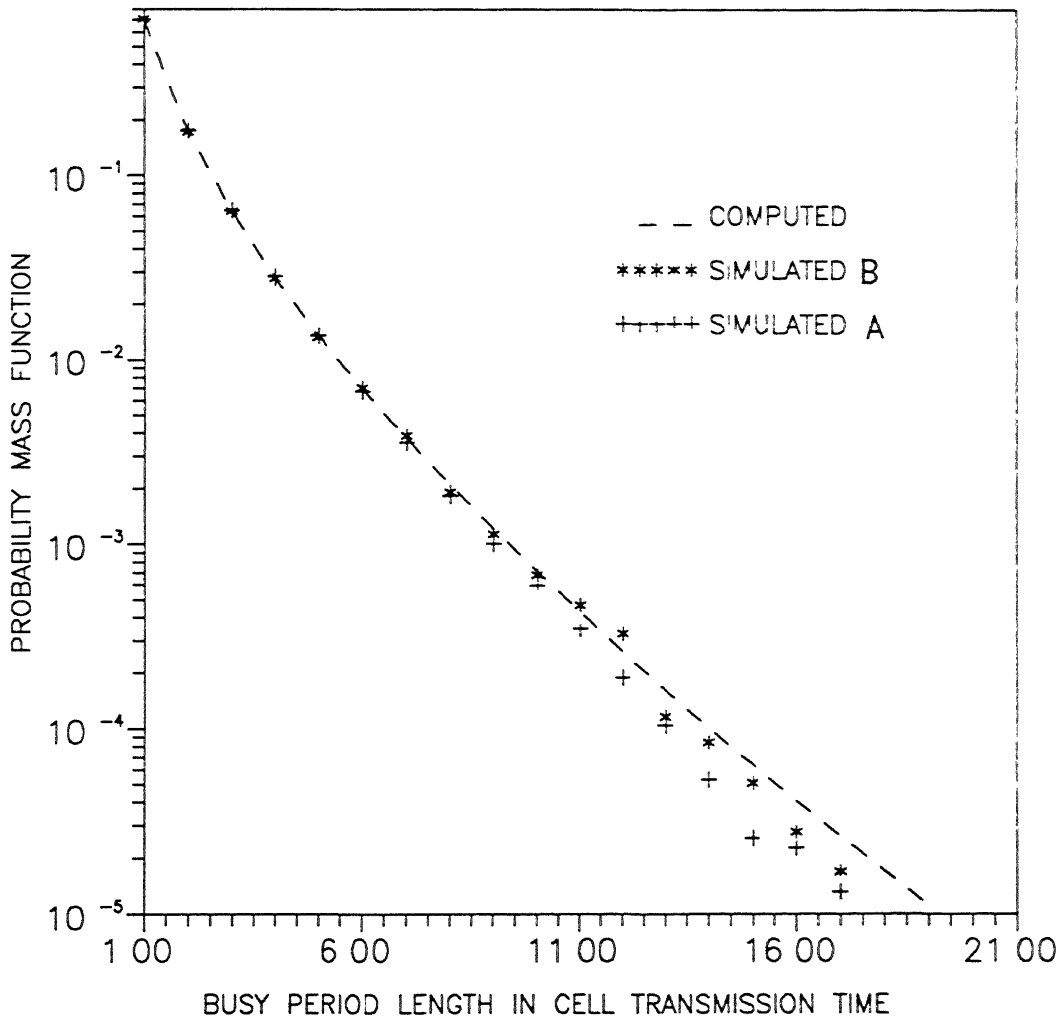


FIG 5 6 P M F OF BUSY PERIOD LENGTH OF MMPP/D/1 QUEUE OBTAINED THROUGH COMPUTATION AND SIMULATION FOR $N_s=150$ ($\rho = 0.35$)

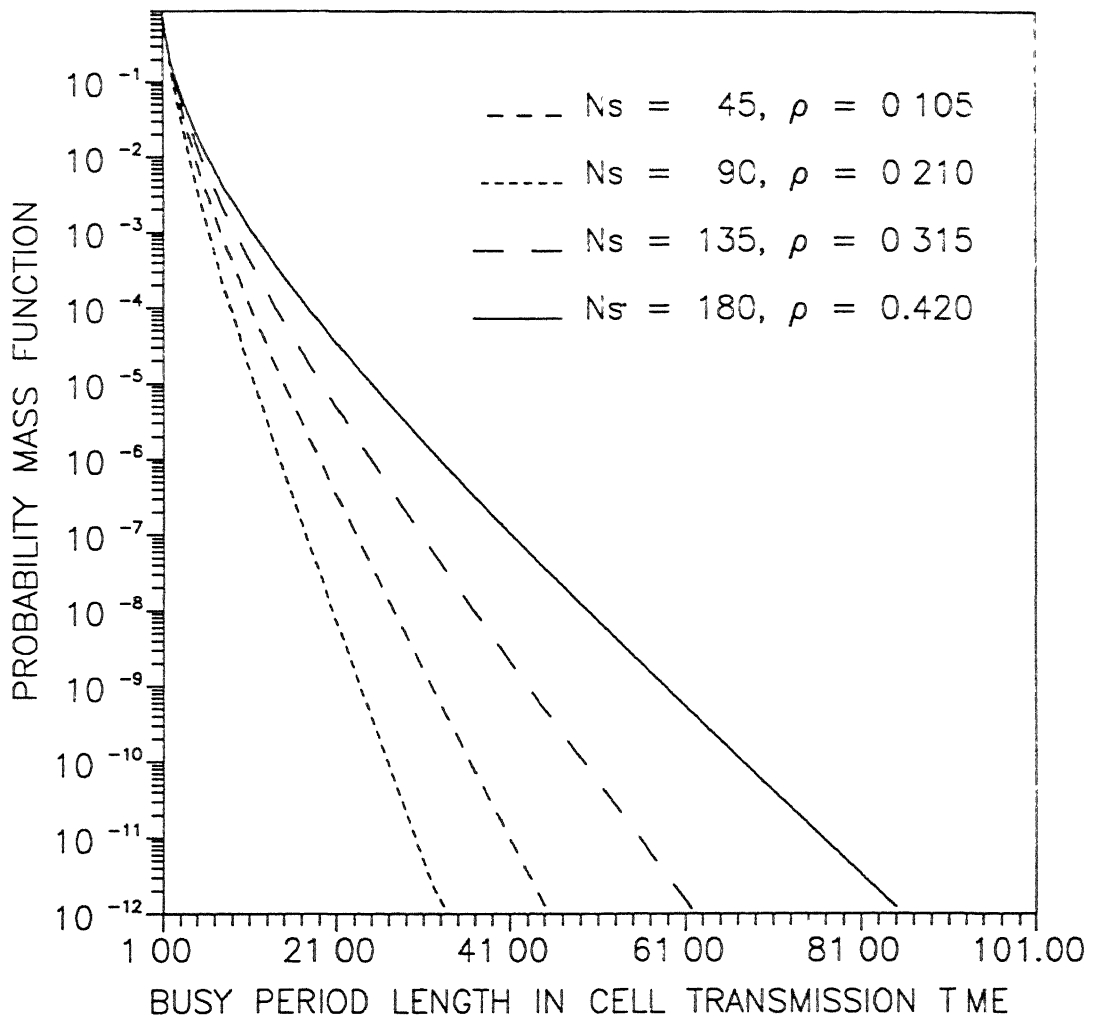


FIG 5.7 PROBABILITY MASS FUNCTION OF THE BUSY PERIOD
 LENGTH OF MMPP/D/1 QUEUE FOR $\rho < 0.5$

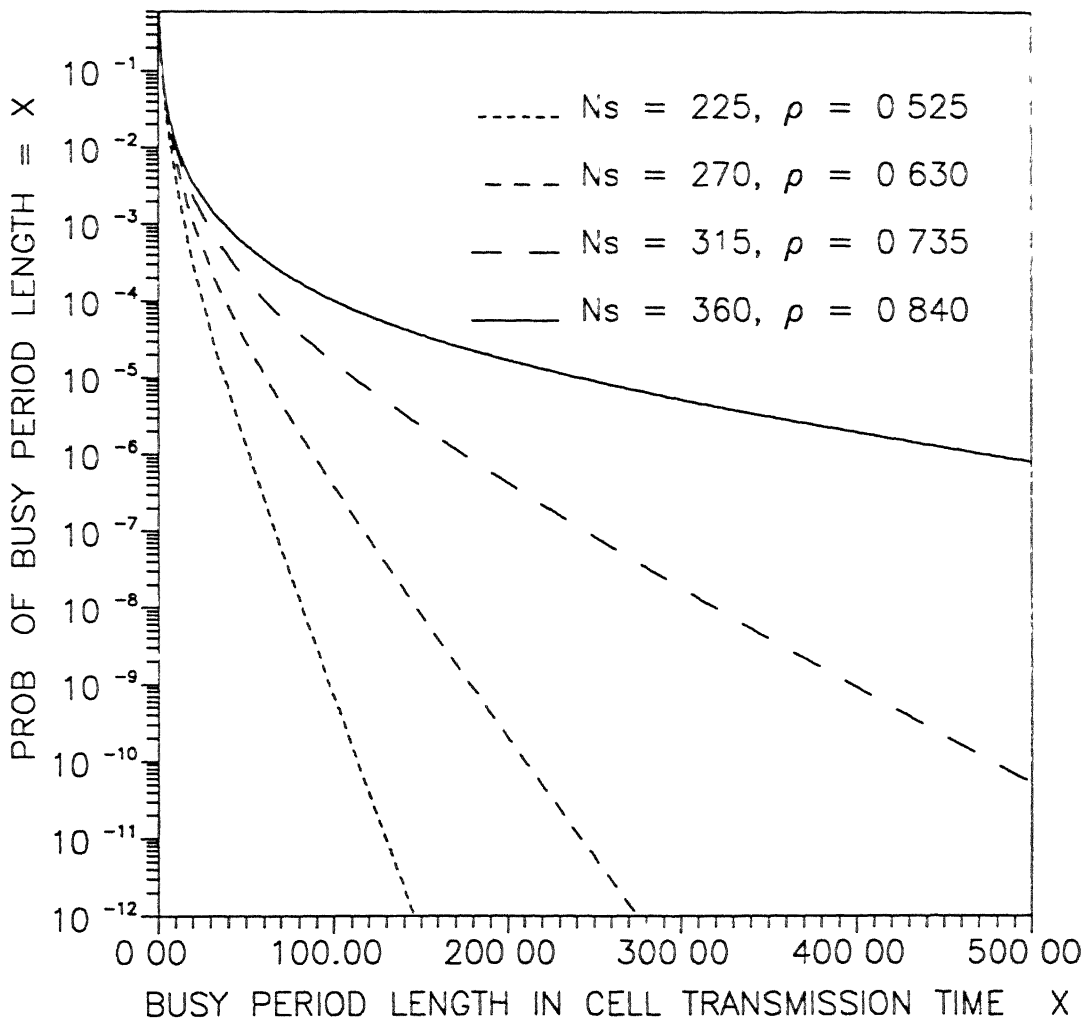


FIG 5.8 PROBABILITY MASS FUNCTION OF THE BUSY PERIOD LENGTH OF A MMPP/D/1 QUEUE COMPUTED FOR $\rho > 0.5$.

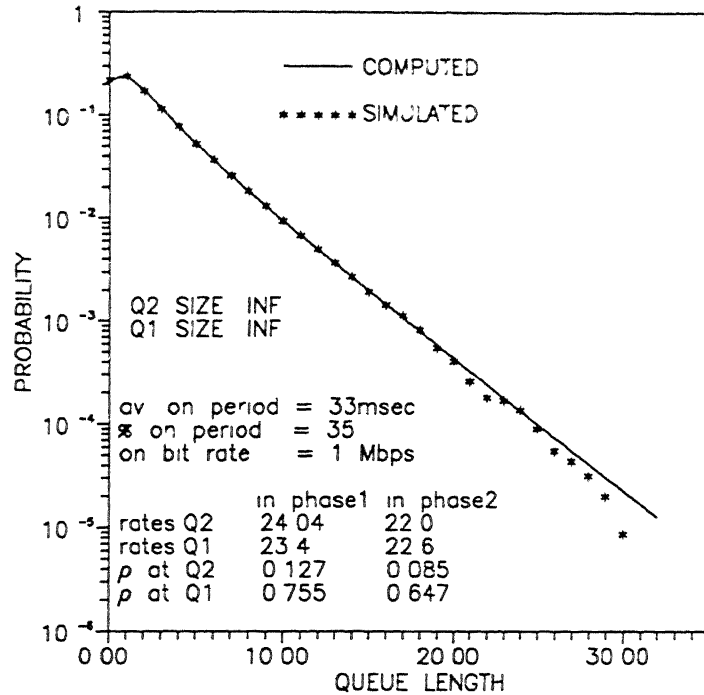


FIG 5.9 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,45) TYPE1 SOURCES ($\rho = 0.7, 0.1$)

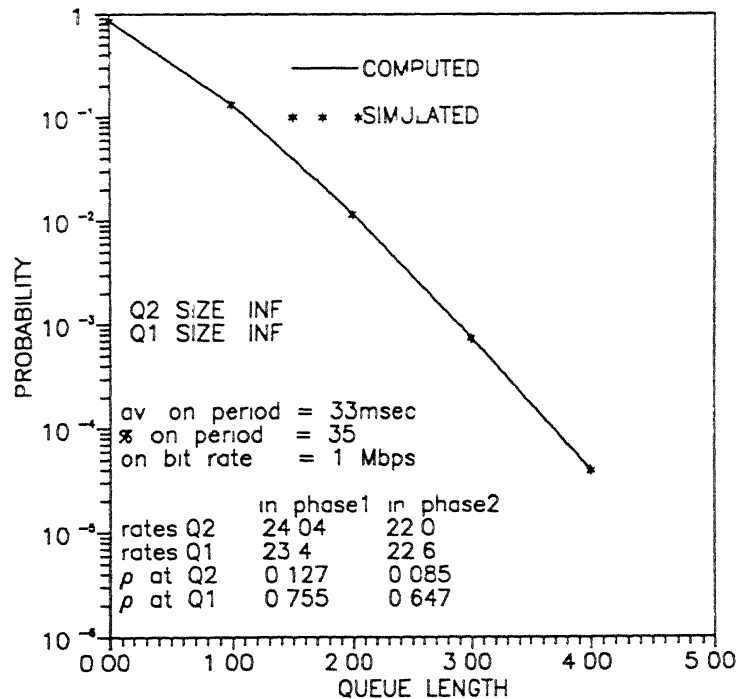


FIG 5.10 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,45) TYPE1 SOURCES ($\rho = 0.7, 0.1$)

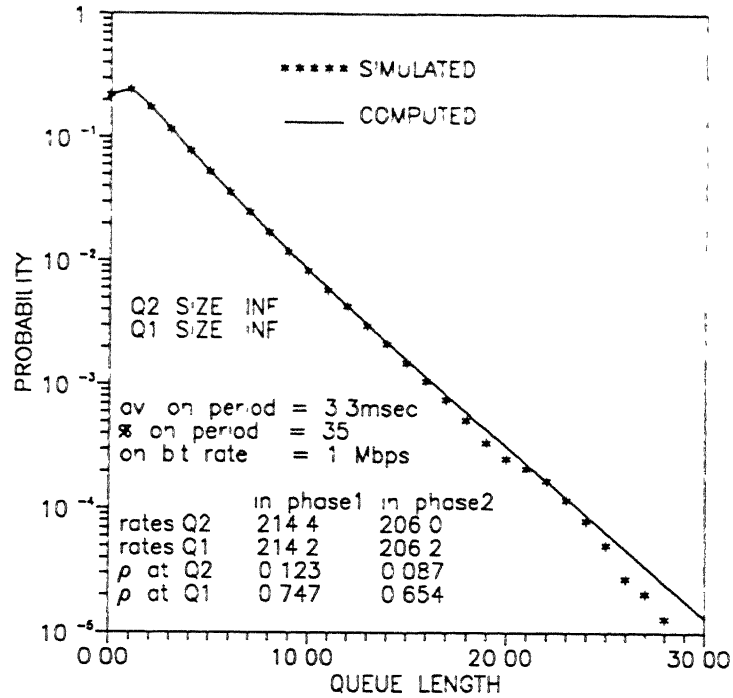


FIG 5.11 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,45) TYPE2 SOURCES ($\rho = 0.7, 0.1$)

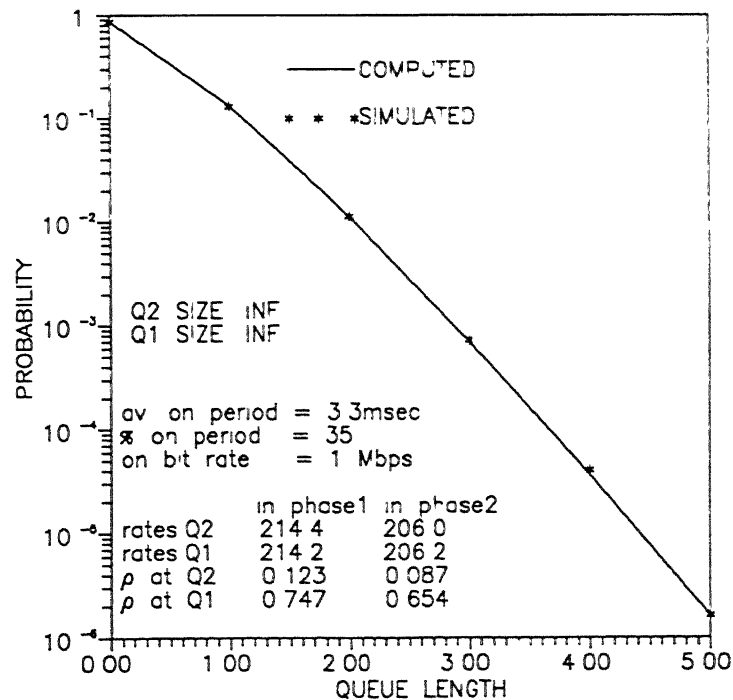


FIG 5.12 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,45) TYPE2 SOURCES ($\rho = 0.7, 0.1$)

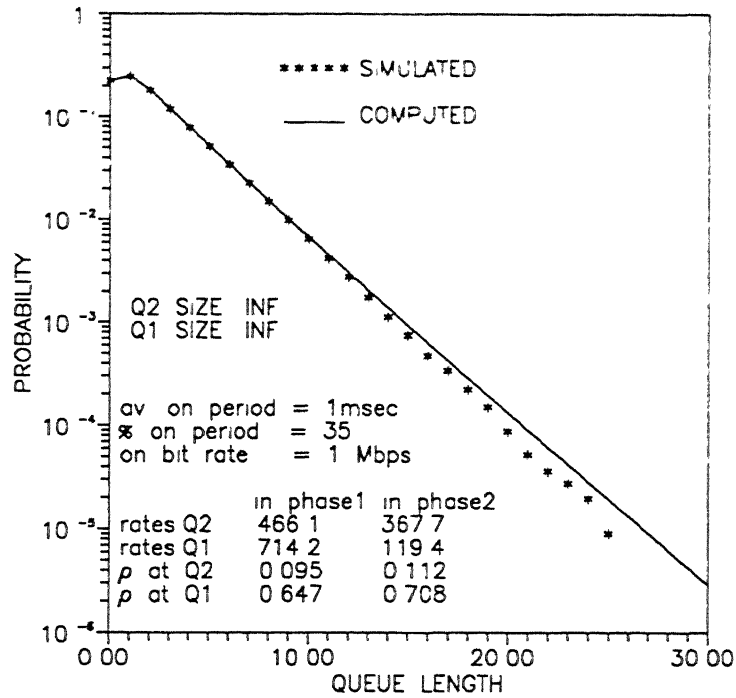


FIG 5.13 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,45) TYPE3 SOURCES ($\rho = 0.7, 0.1$)

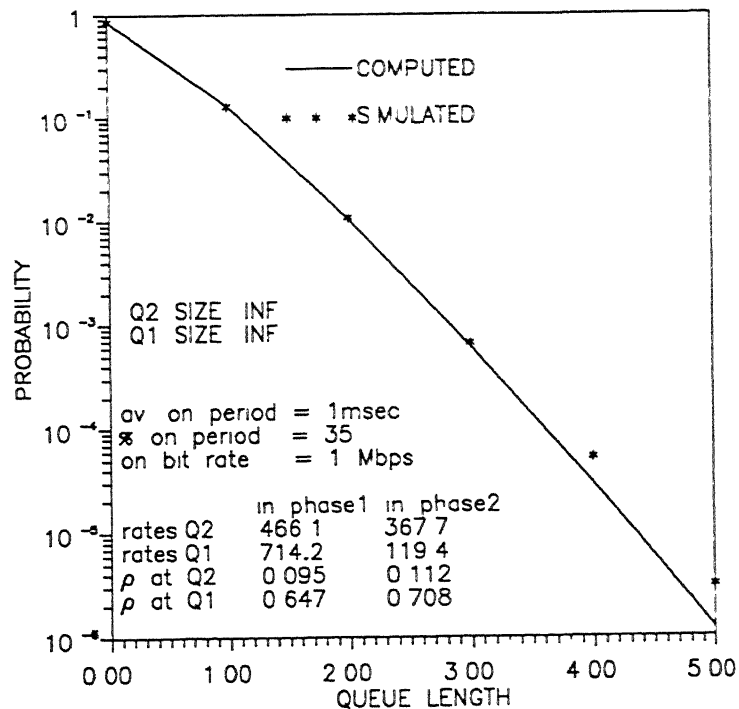


FIG 5.14 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,45) TYPE3 SOURCES ($\rho = 0.7, 0.1$)

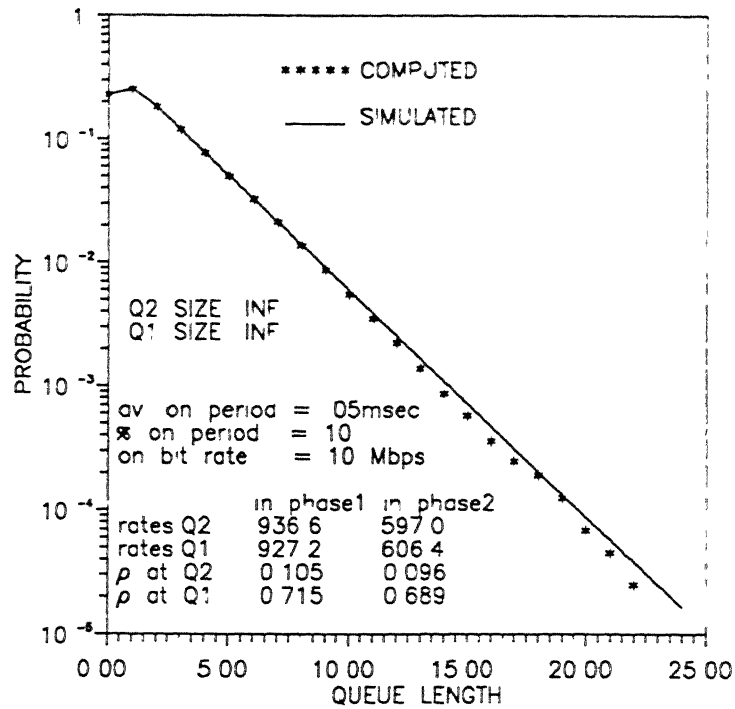


FIG 5.15 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (105,15) TYPE4 SOURCES ($\rho = 0.7, 0.1$)

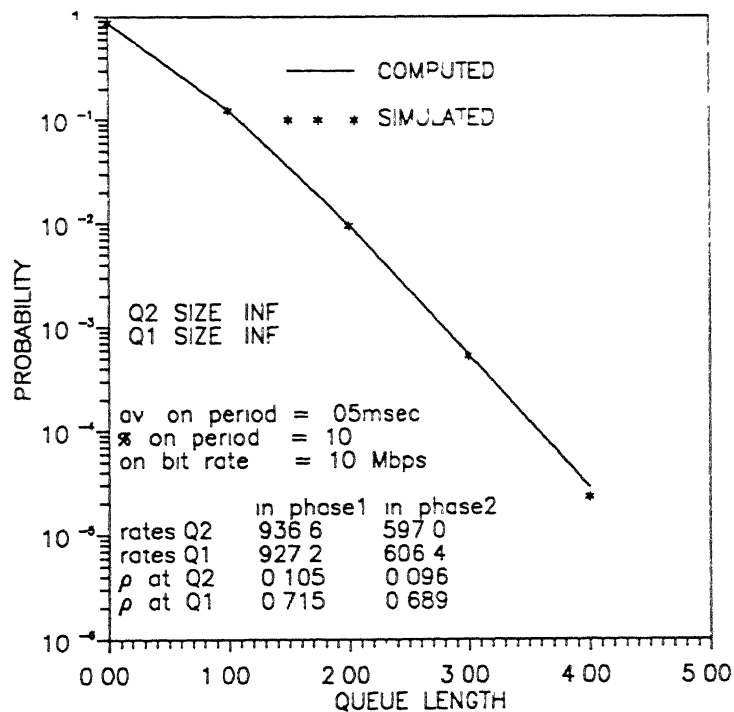


FIG 5.16 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (105,45) TYPE4 SOURCES ($\rho = 0.7, 0.1$)

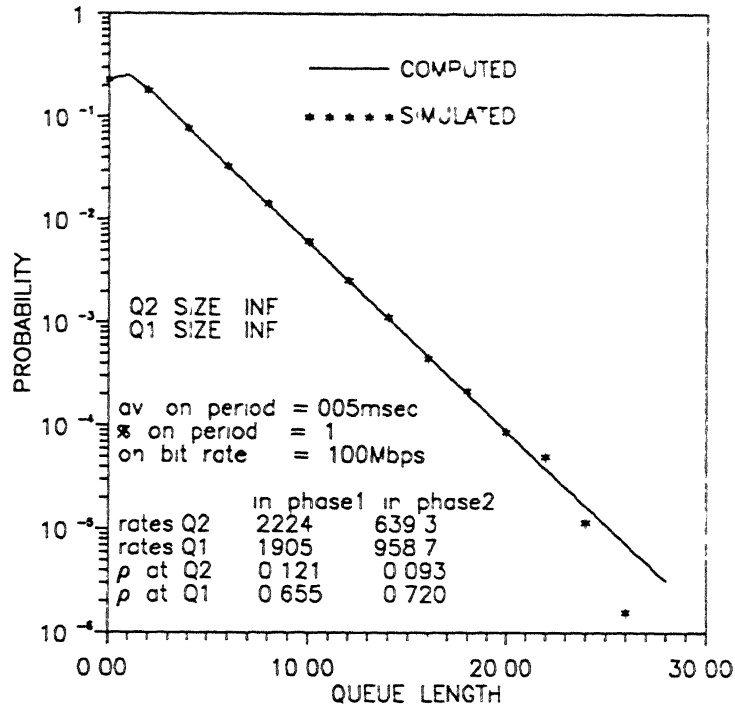


FIG 5.17 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (105,15) TYPE5 SOURCES ($\rho = 0.7, 0.1$)

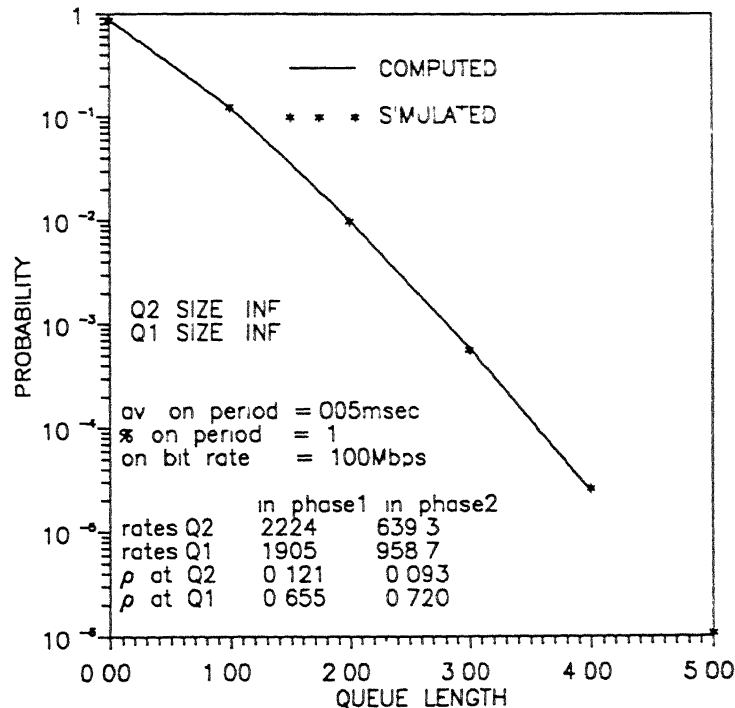


FIG 5.18 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (105,15) TYPE5 SOURCES ($\rho = 0.7, 0.1$)

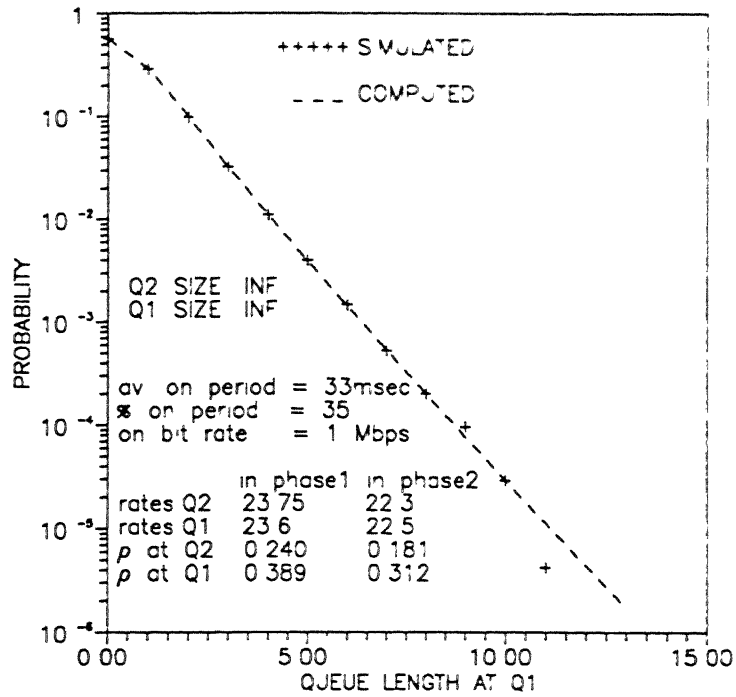


FIG 5.19 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (150,90) TYPE1 SOURCES ($\rho = 0.35, 0.2$)

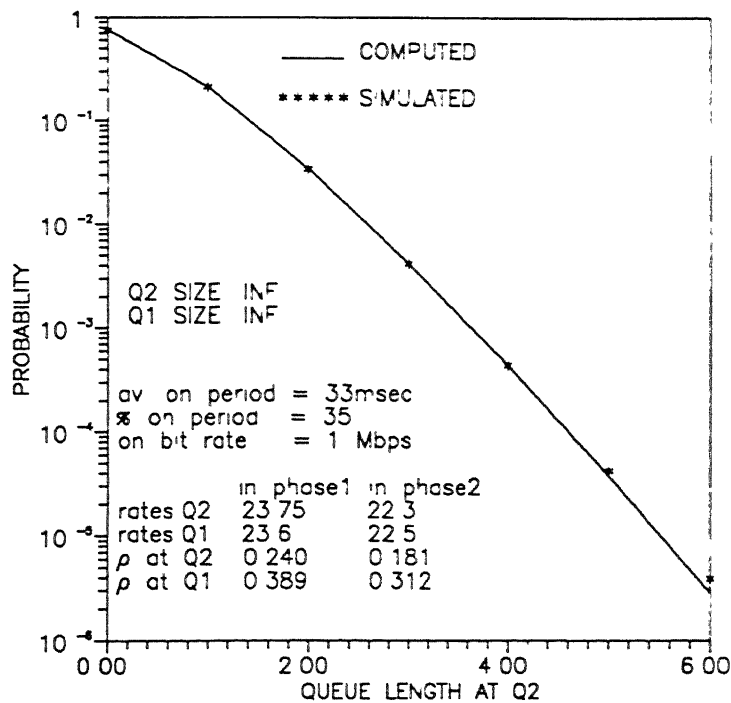


FIG 5.20 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (150,90) TYPE1 SOURCES ($\rho = 0.35, 0.2$)

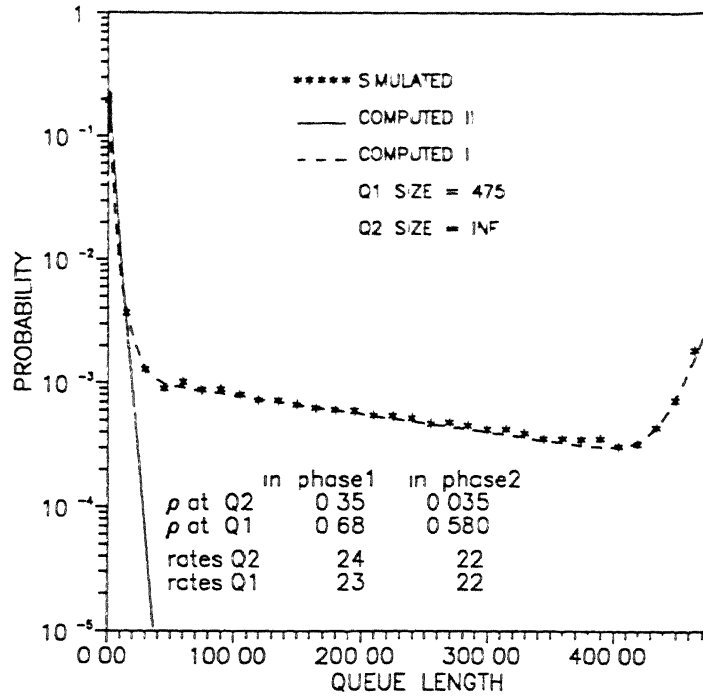


FIG 5.21 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR A (300 TYPE1, BURSTY) TRAFFIC AT Q1, Q2

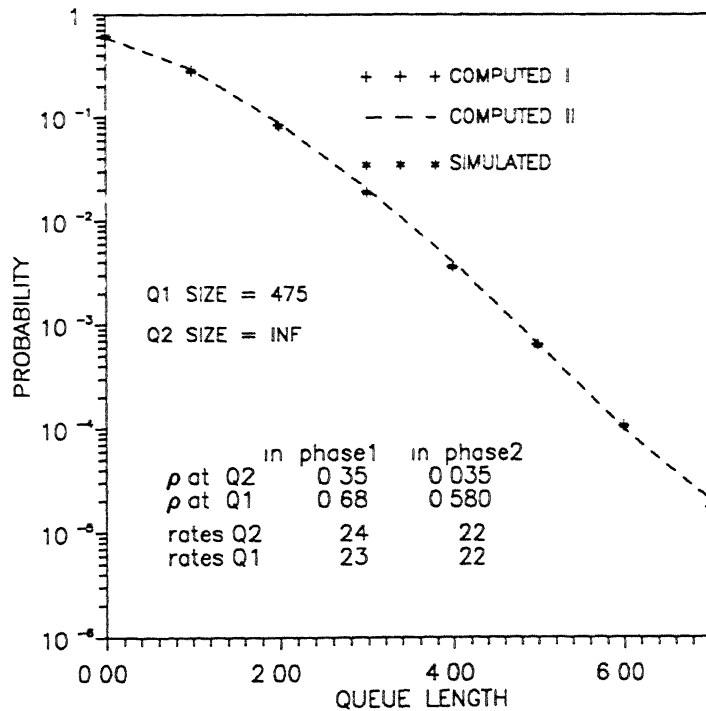


FIG 5.22 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR A (300 TYPE1, BURSTY) TRAFFIC AT Q1, Q2

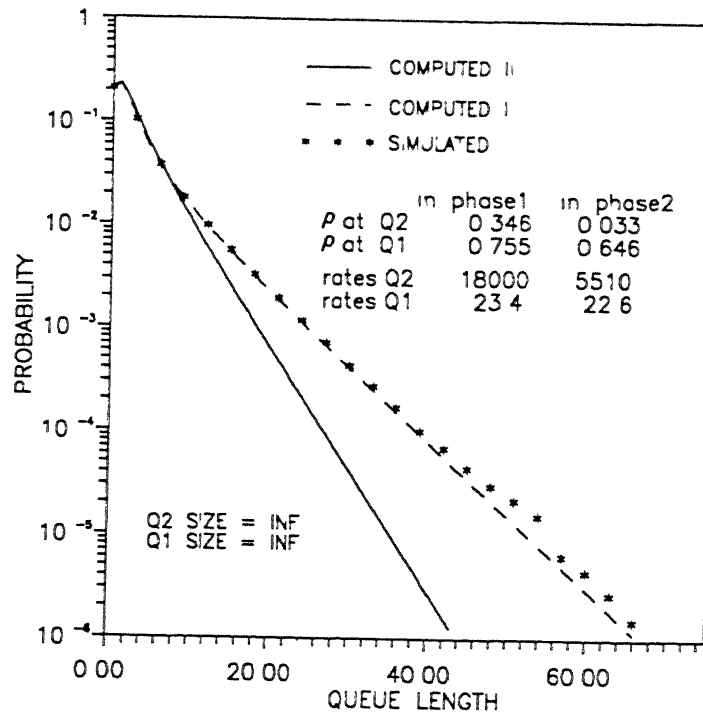


FIG 5.23 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED
FOR (300 TYPE1, 16 TYPE4) SOURCES ($\text{AV } \rho = 0.70, 0.1$)

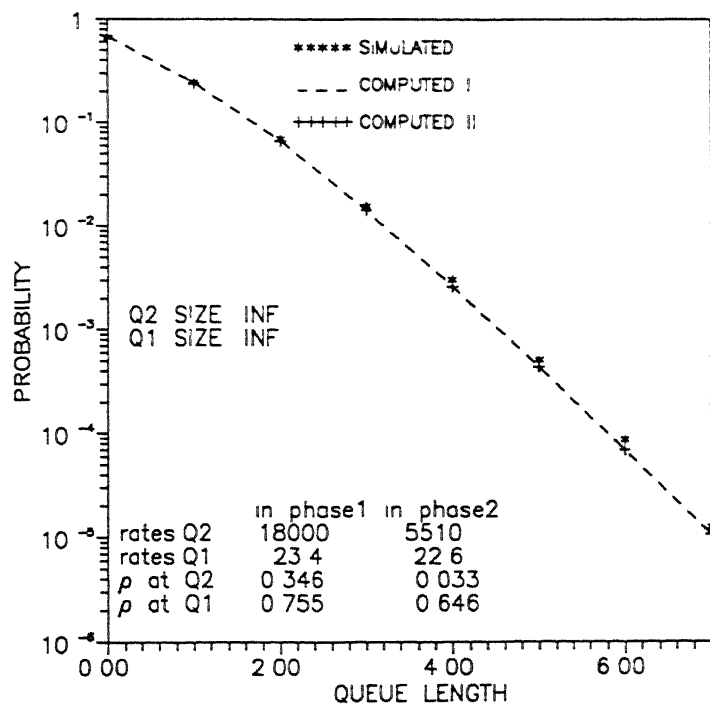


FIG 5.24 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED
FOR (300 TYPE1, 16 TYPE4) SOURCES ($\rho = 0.7, 0.1$)

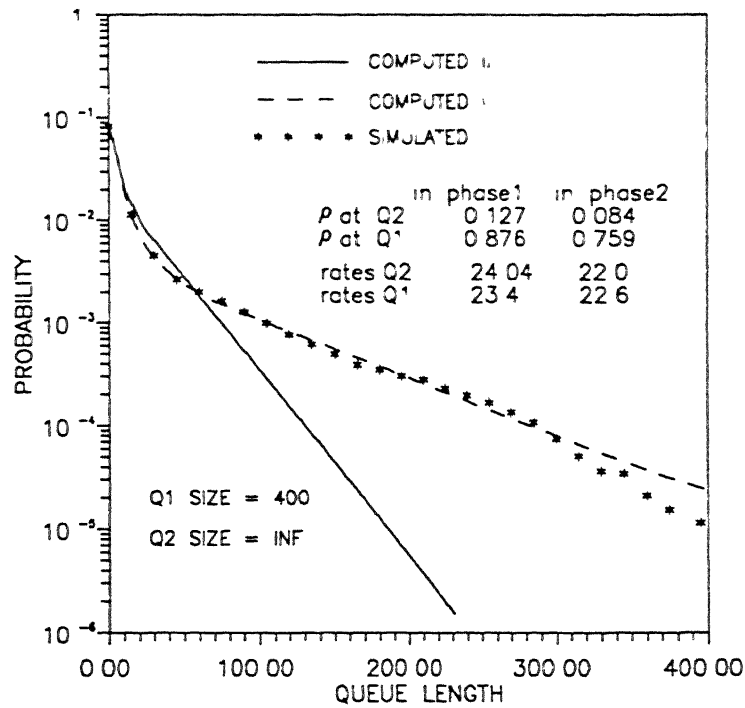


FIG 5.25 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (350,45) TYPE1 SOURCES AND Q1 SIZE=400 ($\rho = 0.81, 0.1$)

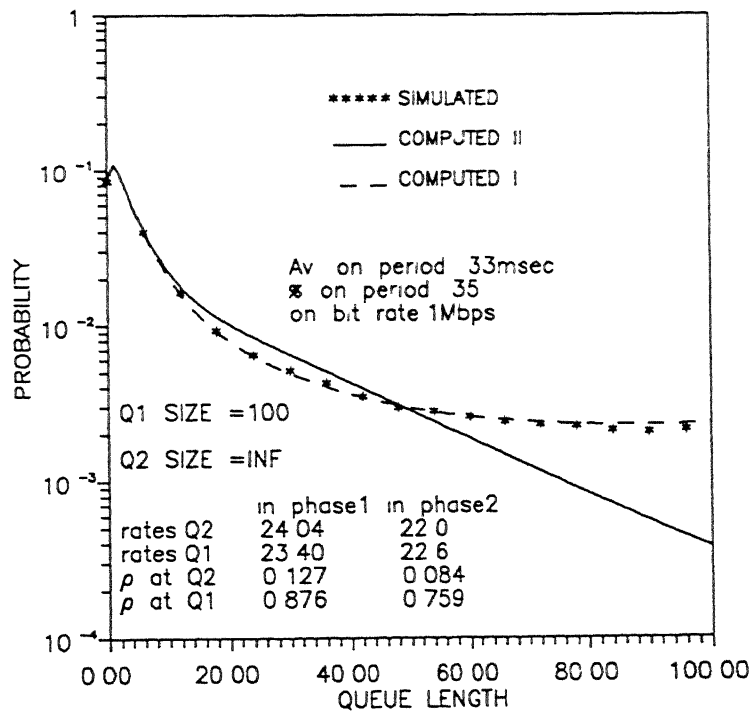


FIG 5.26 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (350,45) TYPE1 SOURCES AND Q1 SIZE=100 ($\rho = 0.81, 0.1$)

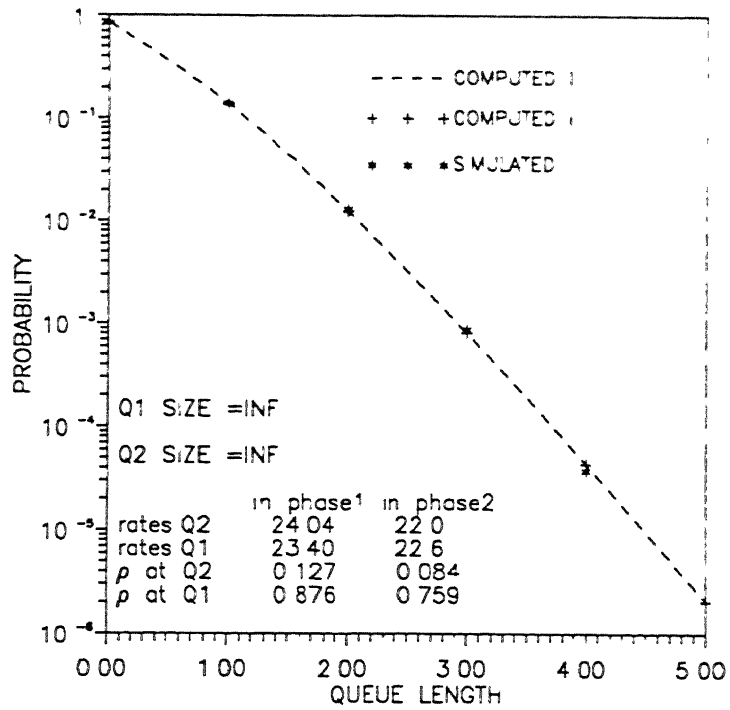


FIG 5.27 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (350,45) TYPE1 SOURCES AND Q1 SIZE=INF ($\rho = 0.81, 0.1$)

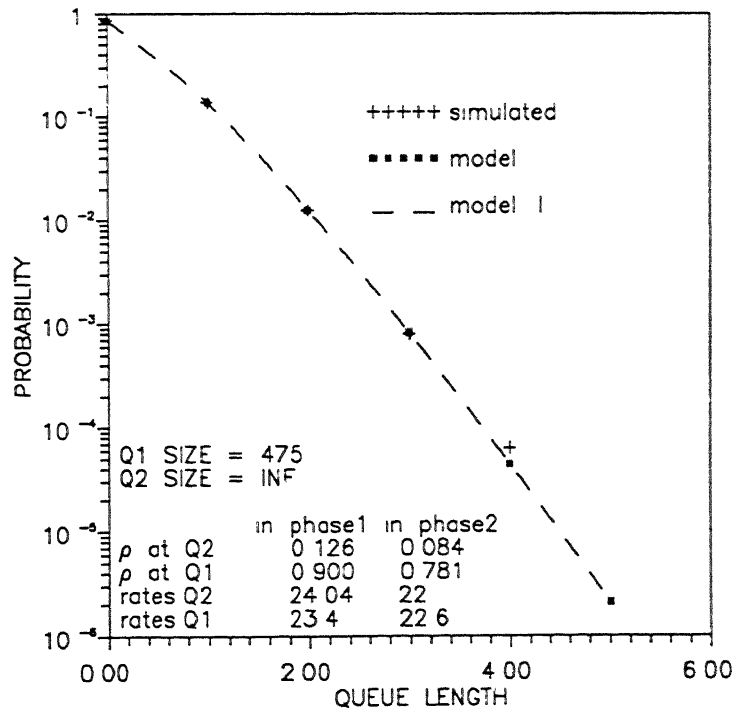


FIG 5.28 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (360,45) TYPE1 SOURCES AND Q1 SIZE = 475 ($\rho = 0.84, 0.1$)

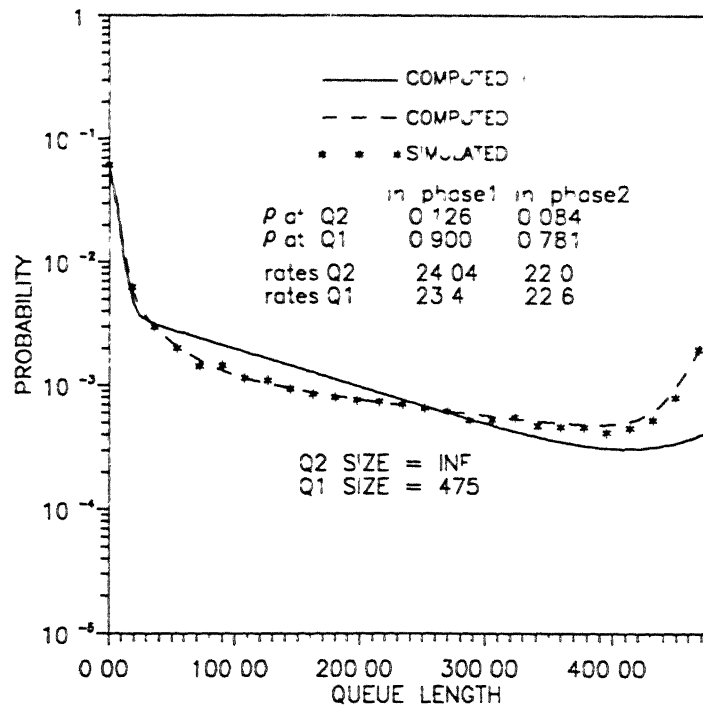


FIG 5.29 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (360,45) TYPE1 SOURCES ($\rho = 0.84, 0.1$)

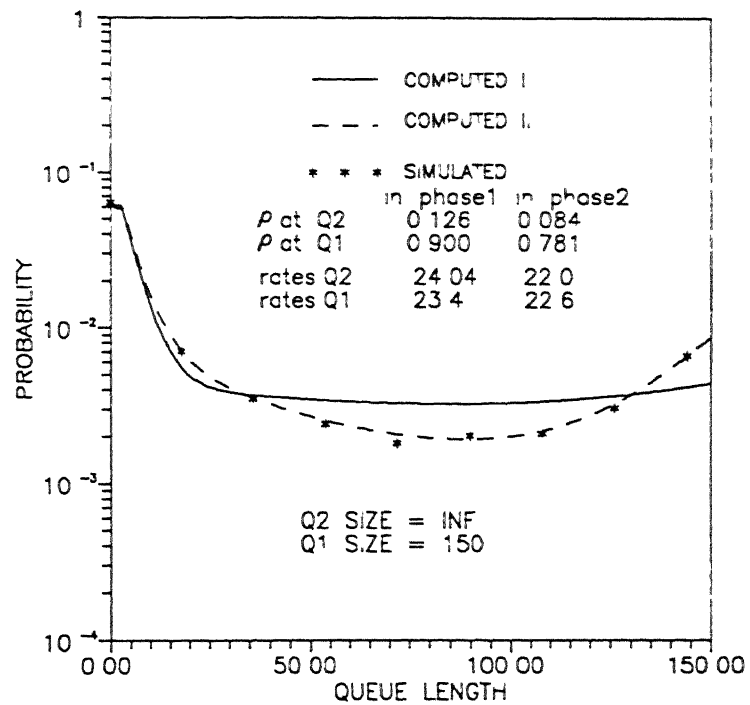


FIG 5.30 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (360,45) TYPE1 SOURCES ($\rho = 0.84, 0.1$)

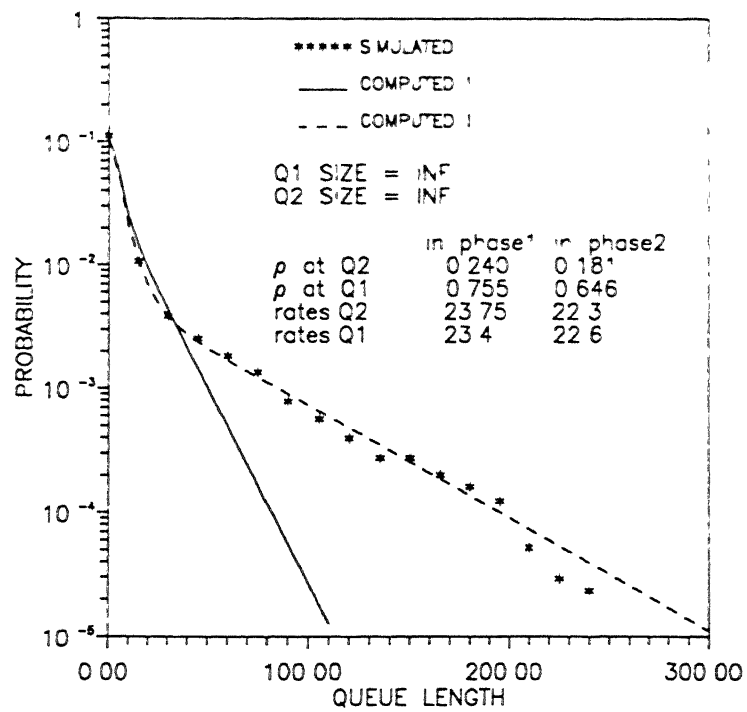


FIG 5.31 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE1 SOURCES AND Q1 SIZE = INF ($\rho = 0.70, 0.2$)

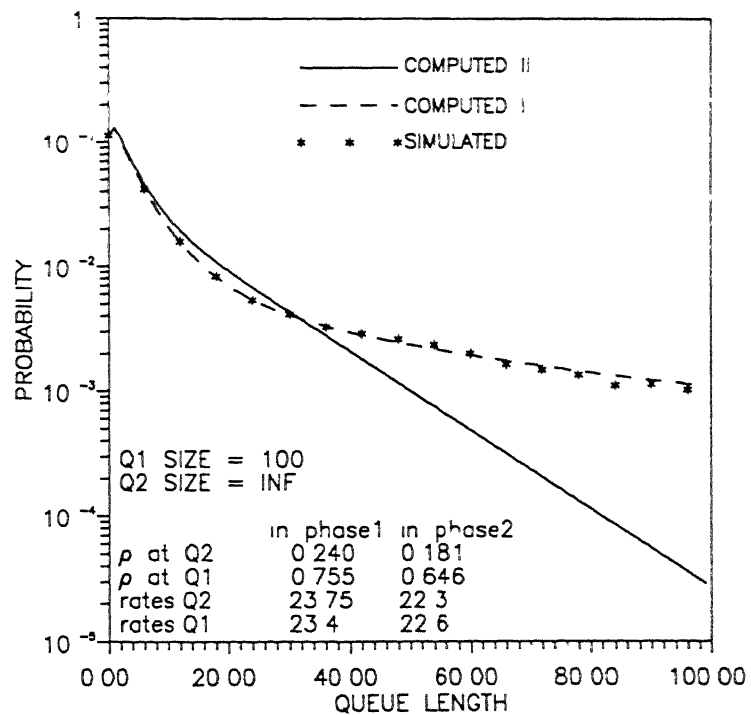


FIG 5.32 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE1 SOURCES AND Q1 SIZE = 100 ($\rho = 0.70, 0.2$)

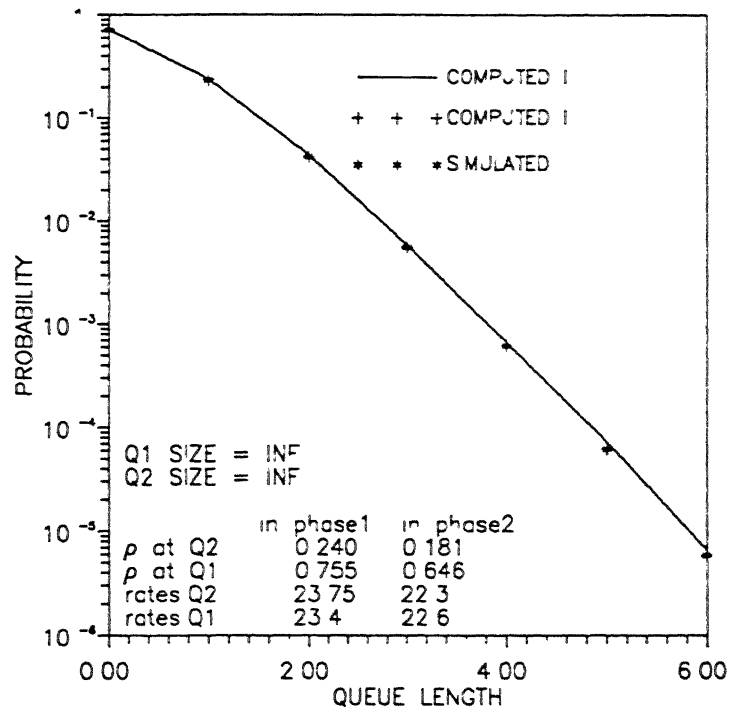


FIG 5.33 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE1 SOURCES AND Q1 SIZE = INF ($\rho = 0.7, 0.1$)

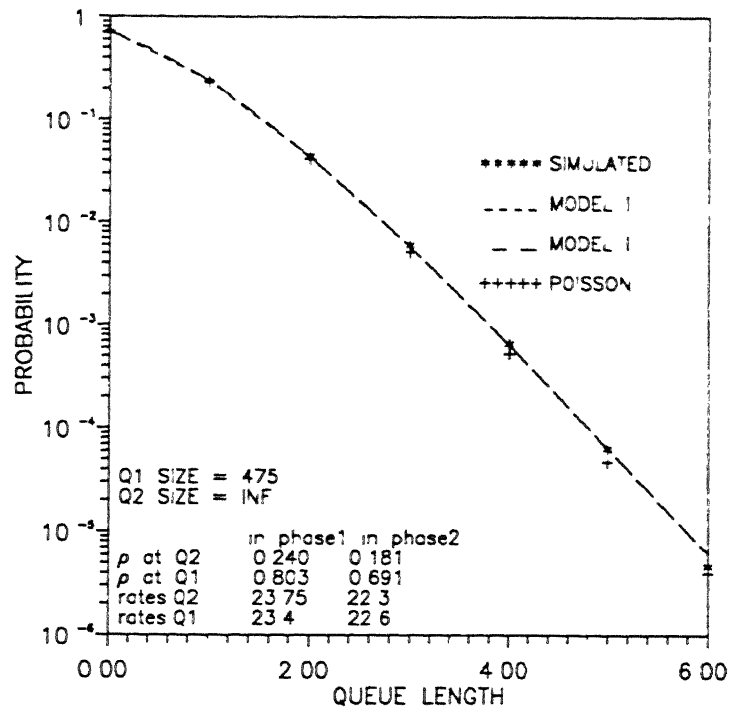


FIG 5.34 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (320,90) TYPE1 SOURCES AND Q1 SIZE = 475 ($\rho = 0.74, 0.1$)

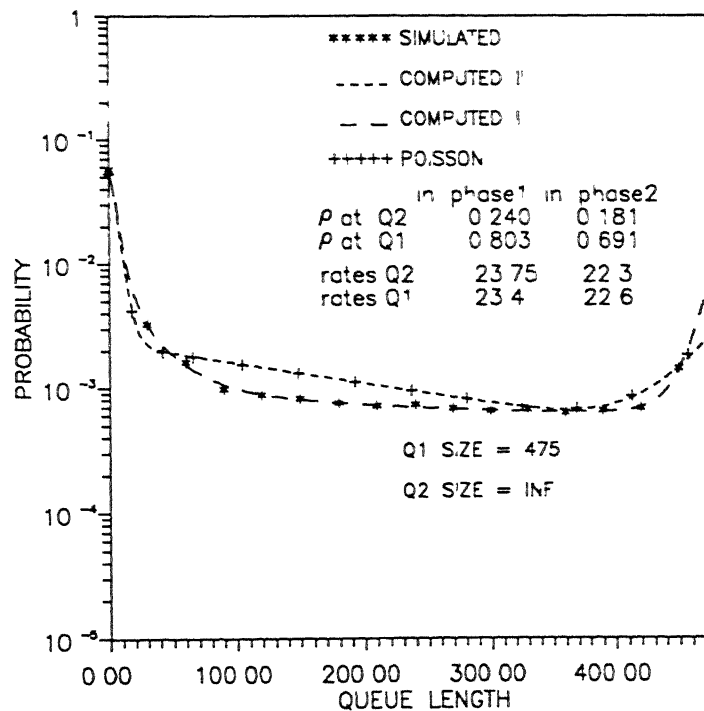


FIG 5.35 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (320,90) TYPE1 SOURCES AND Q1 SIZE = 475 ($\rho = 0.75, 0.2$)

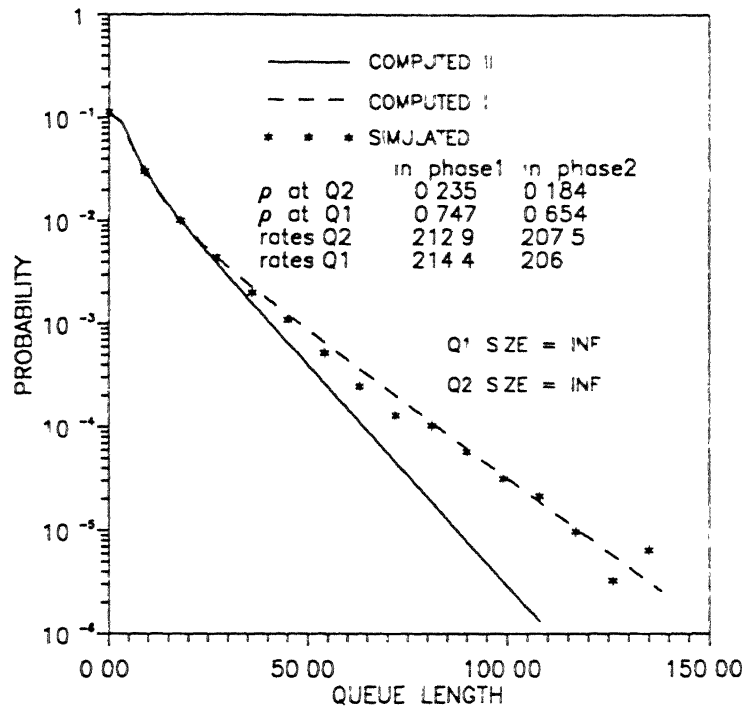


FIG 5.36 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE2 SOURCES AND Q1 SIZE = INF ($\rho = 0.70, 0.2$)

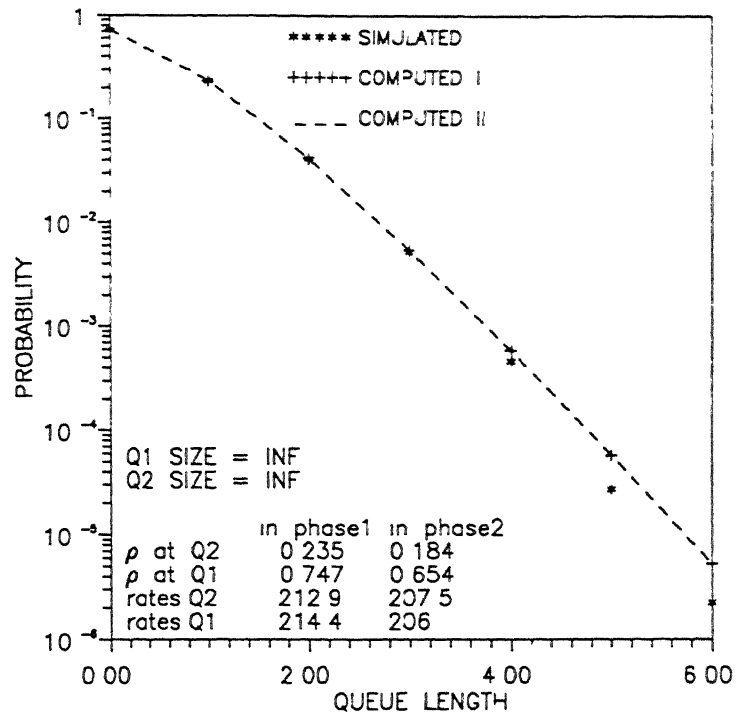


FIG 5.37 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE2 SOURCES AND Q1 SIZE = INF ($\rho = 0.56, 0.31$)

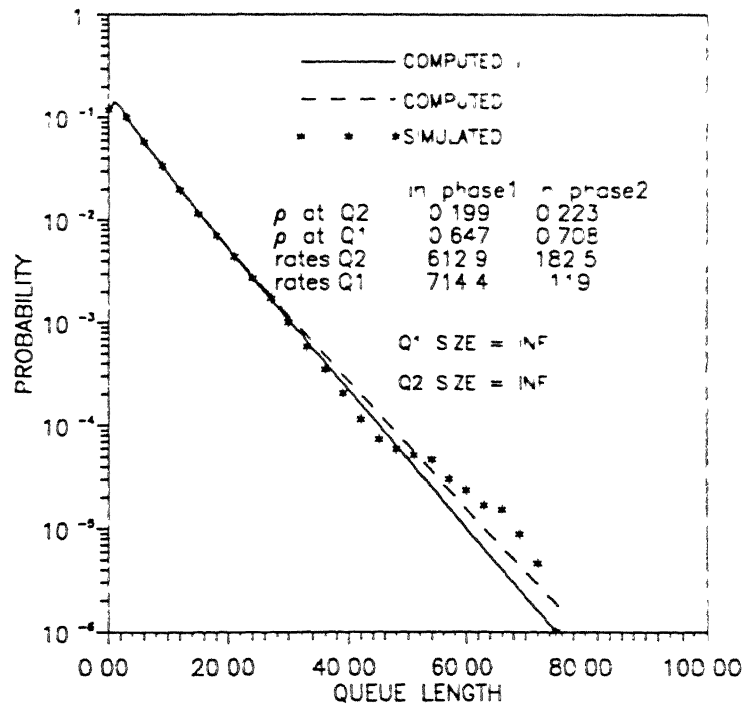


FIG 5.38 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE3 SOURCES AND Q1 SIZE = INF ($\rho = 0.70, 0.2$)

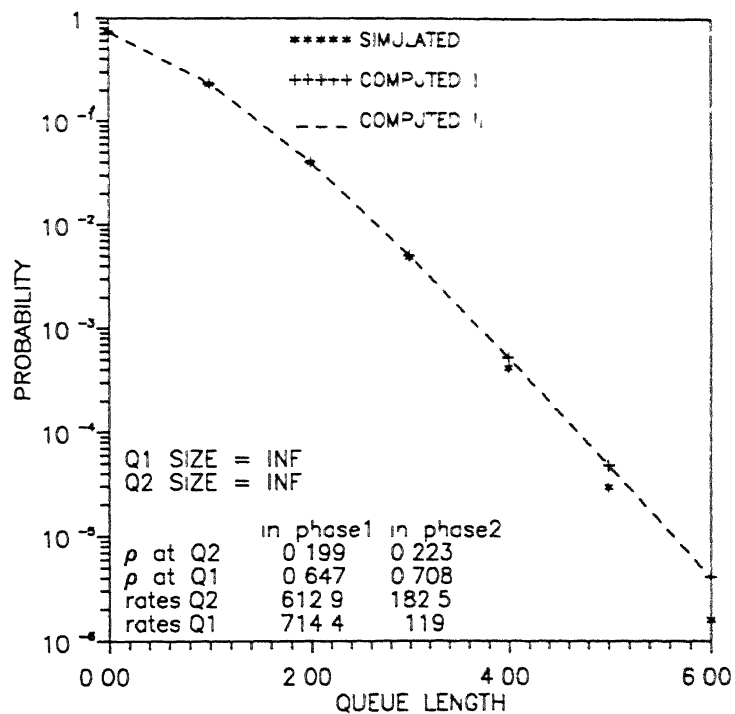


FIG 5.39 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (300,90) TYPE3 SOURCES AND Q1 SIZE = INF ($\rho = 0.7, 0.21$)

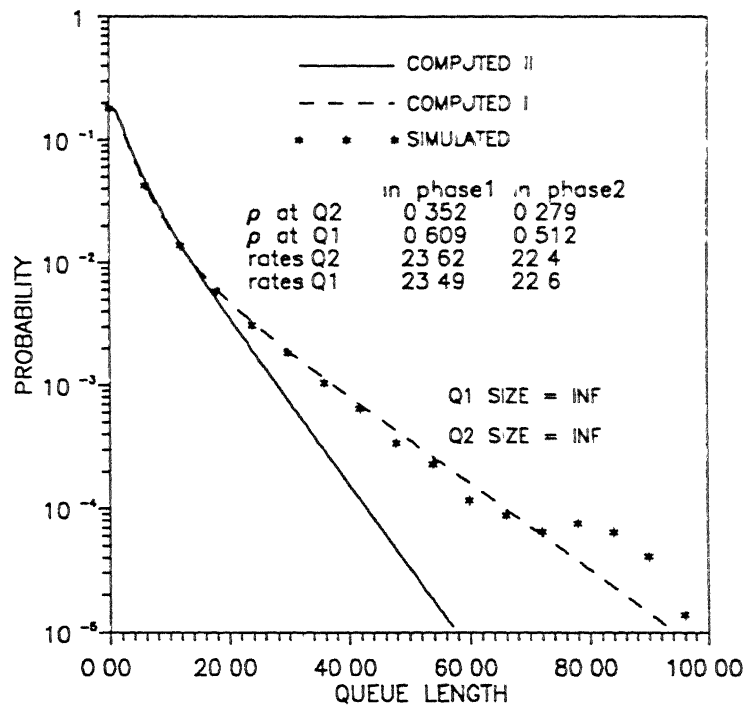


FIG 5.40 QLD OF THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (240,135) TYPE1 SOURCES AND Q1 SIZE = INF ($\rho = 0.56, 0.3$)

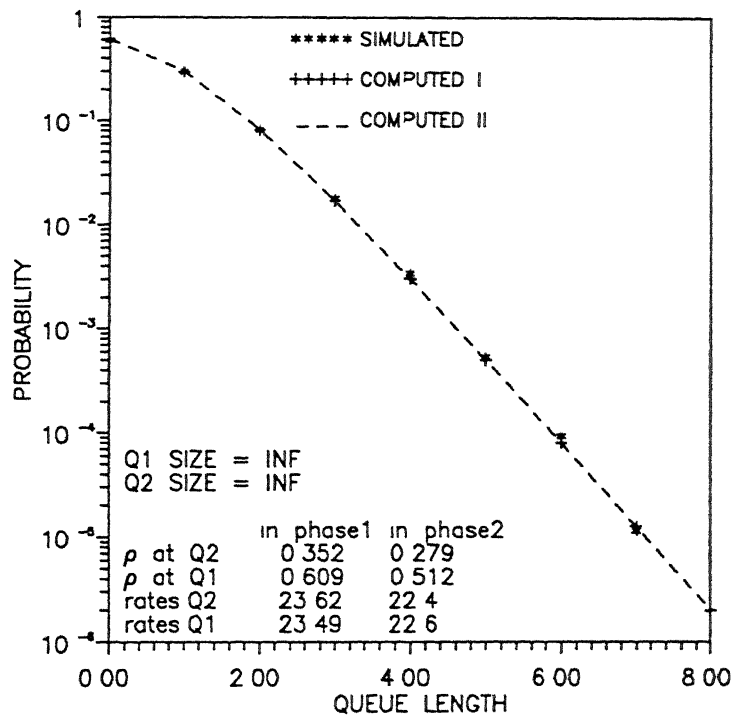


FIG 5.41 QLD OF THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (240,135) TYPE1 SOURCES AND Q1 SIZE = INF ($\rho = 0.56, 0.31$)

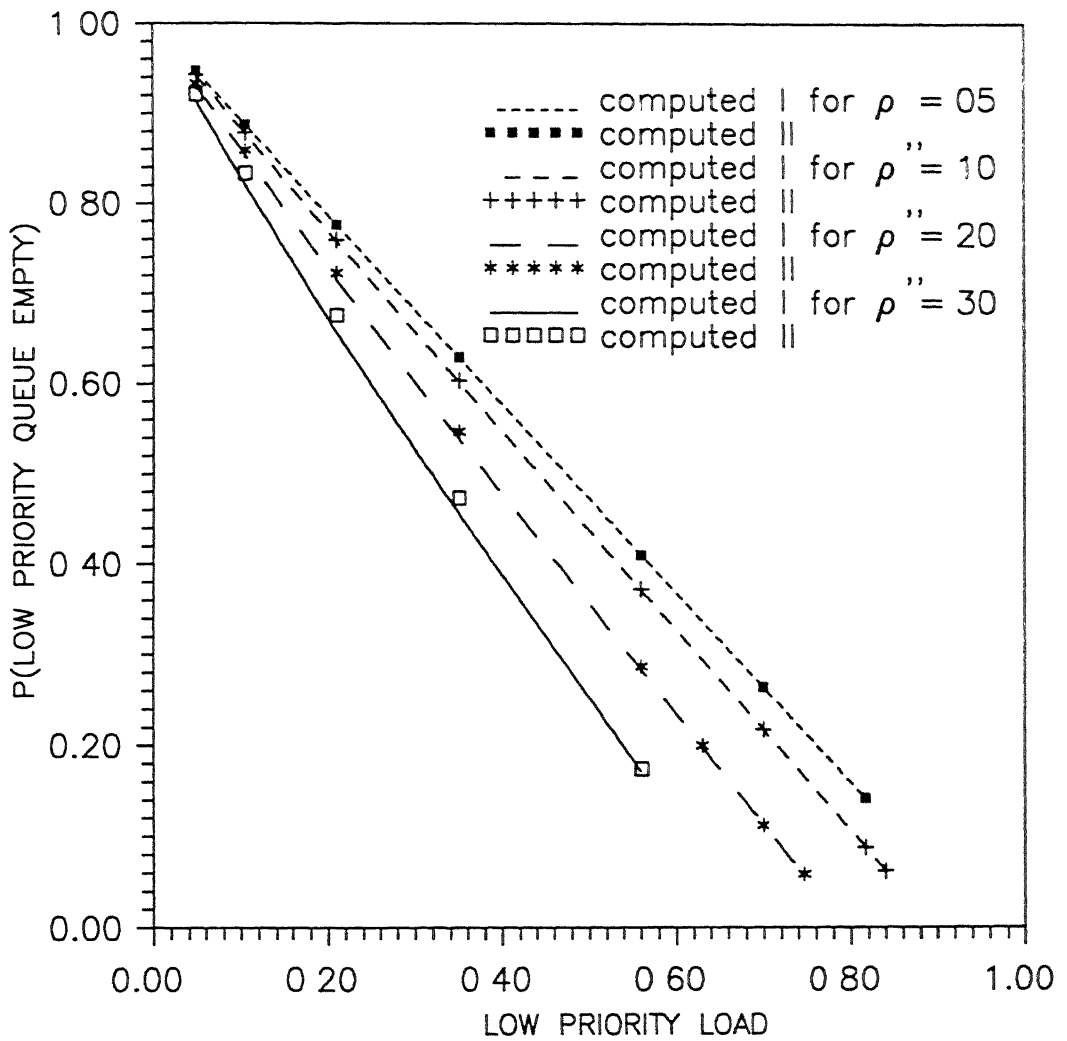


FIG 5.42 VARIATION OF $P(\text{LOW PRIORITY QUEUE EMPTY})$ WITH TRAFFIC OFFERED IN Q1 AND Q2 IN AN MMPP/D/1 QUEUE

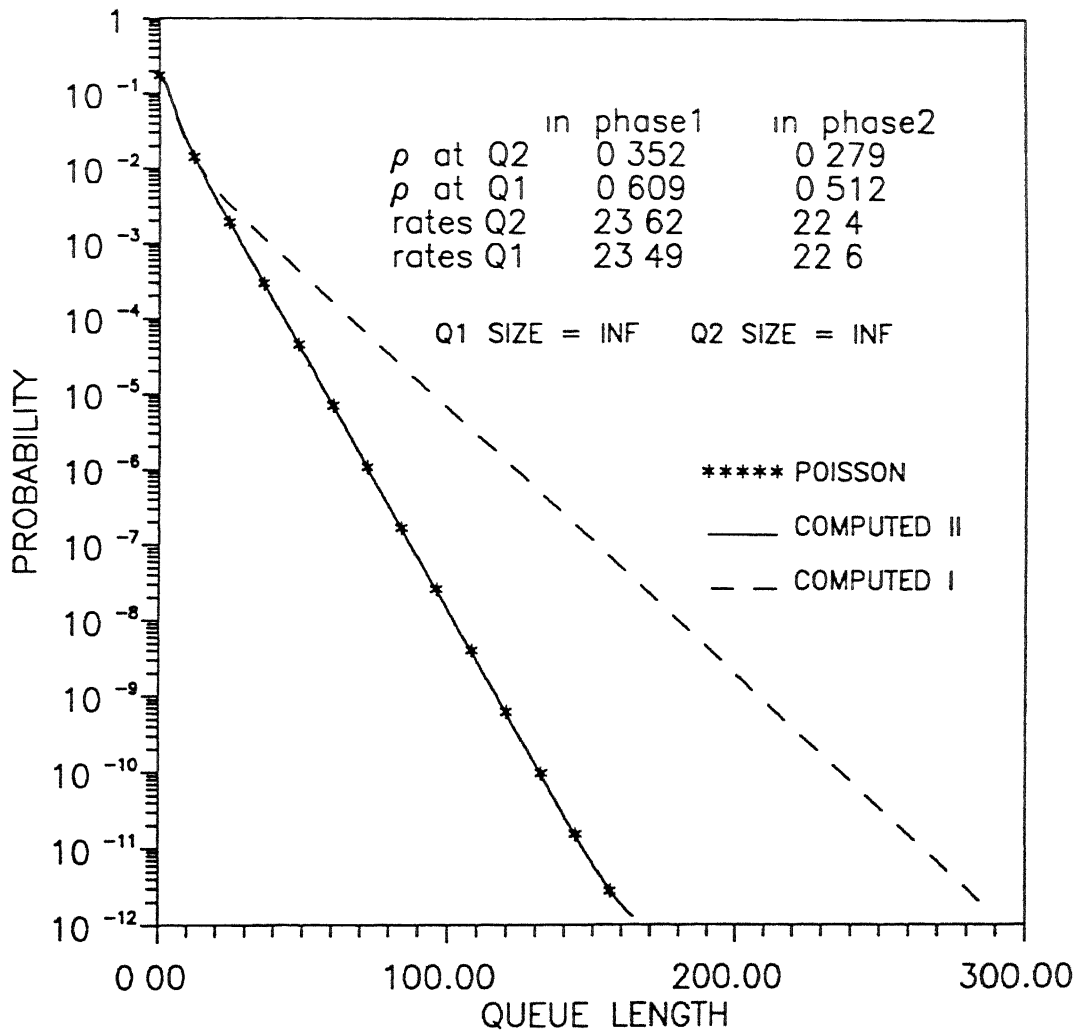


FIG.5.43 QLD OF THE LOW PRIORITY QUEUE COMPUTED USING MODELS I, II AND POISSON FOR (240,135) TYPE1 SOURCES ($\rho = 0.56, 0.31$)

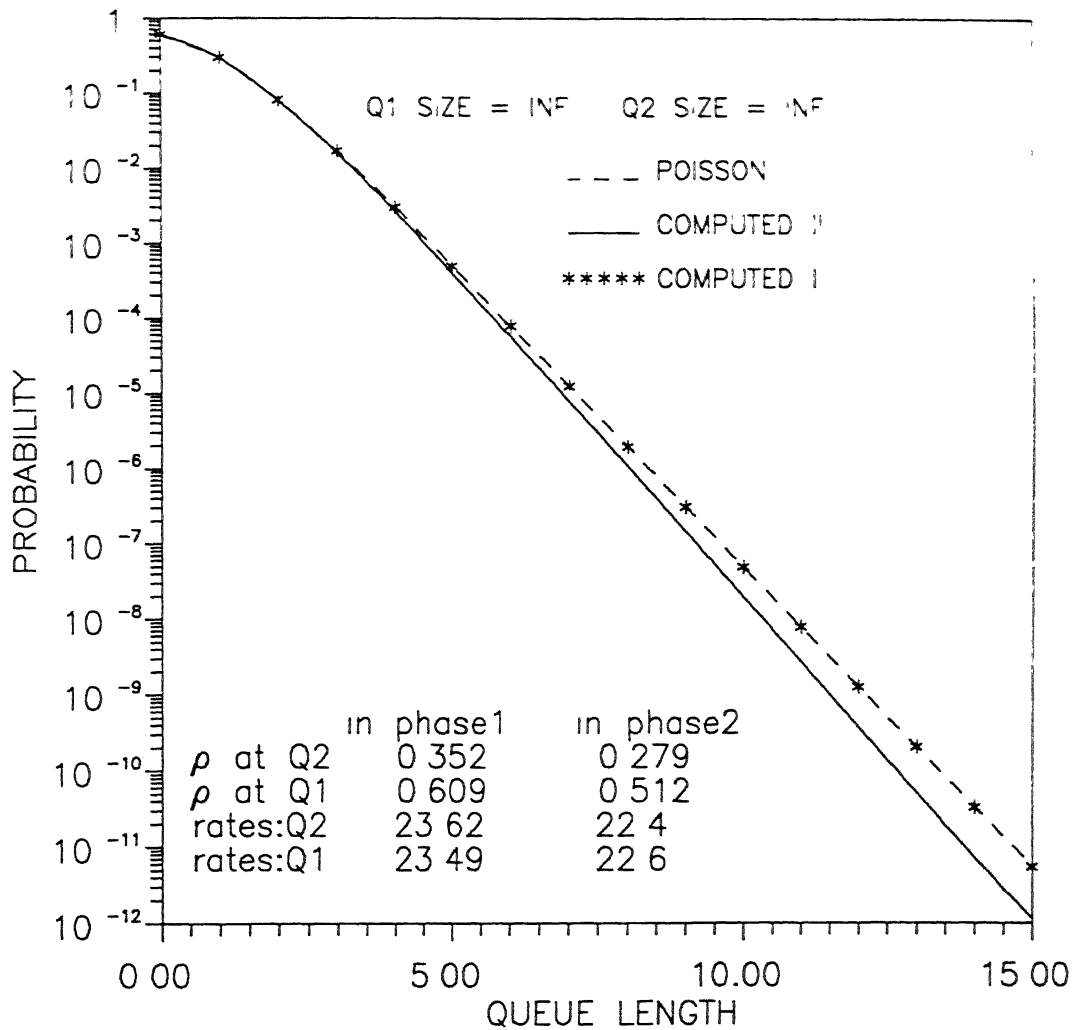


FIG 5.44 QLD OF THE HIGH PRIORITY QUEUE COMPUTED USING MODELS I, II AND POISSON FOR (240,135) TYPE1 SOURCES ($\rho = 0.56, 0.31$)

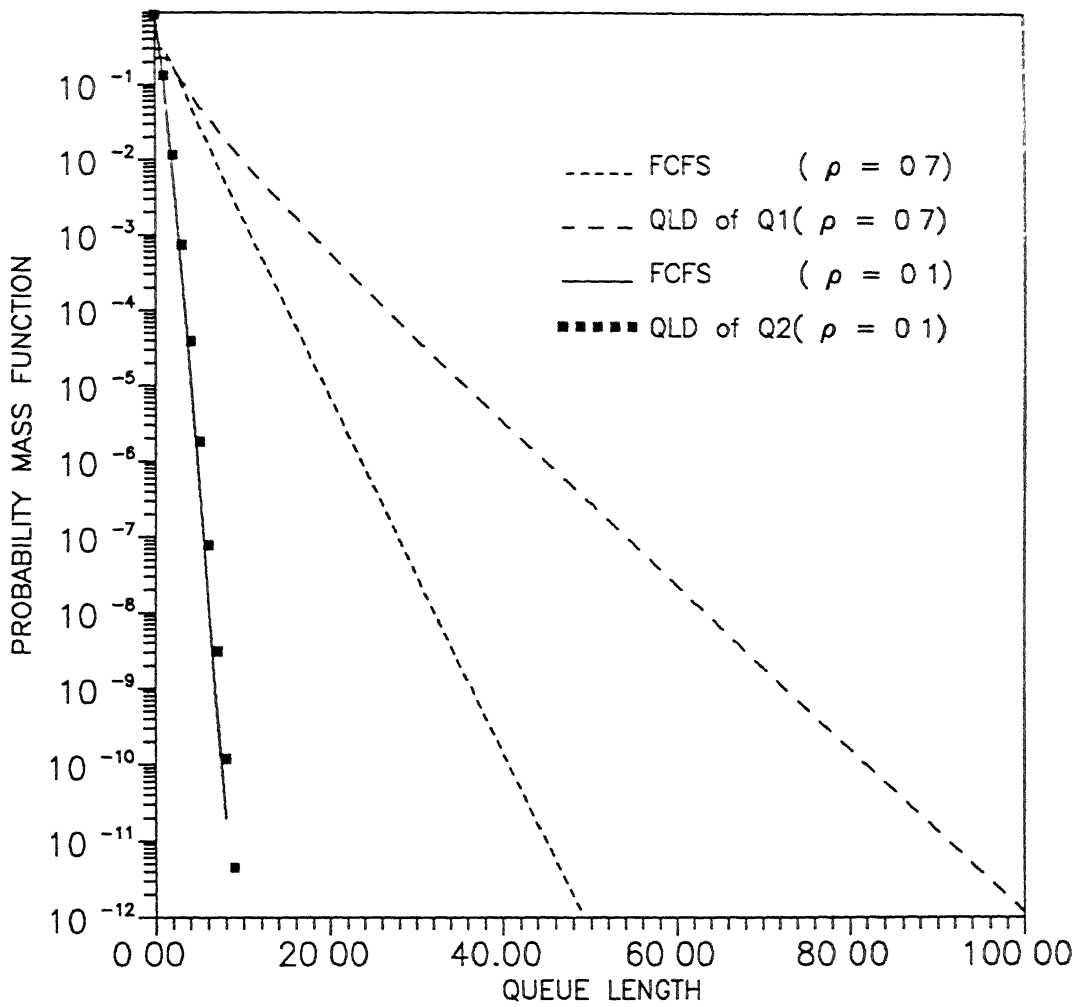


FIG 5.45 COMPARISON OF THE QLDs OF Q1 AND Q2 WITH FCFS QUEUES WITH DEDICATED SERVERS FOR TRAFFICS FROM (300,45) TYPE1 SOURCES

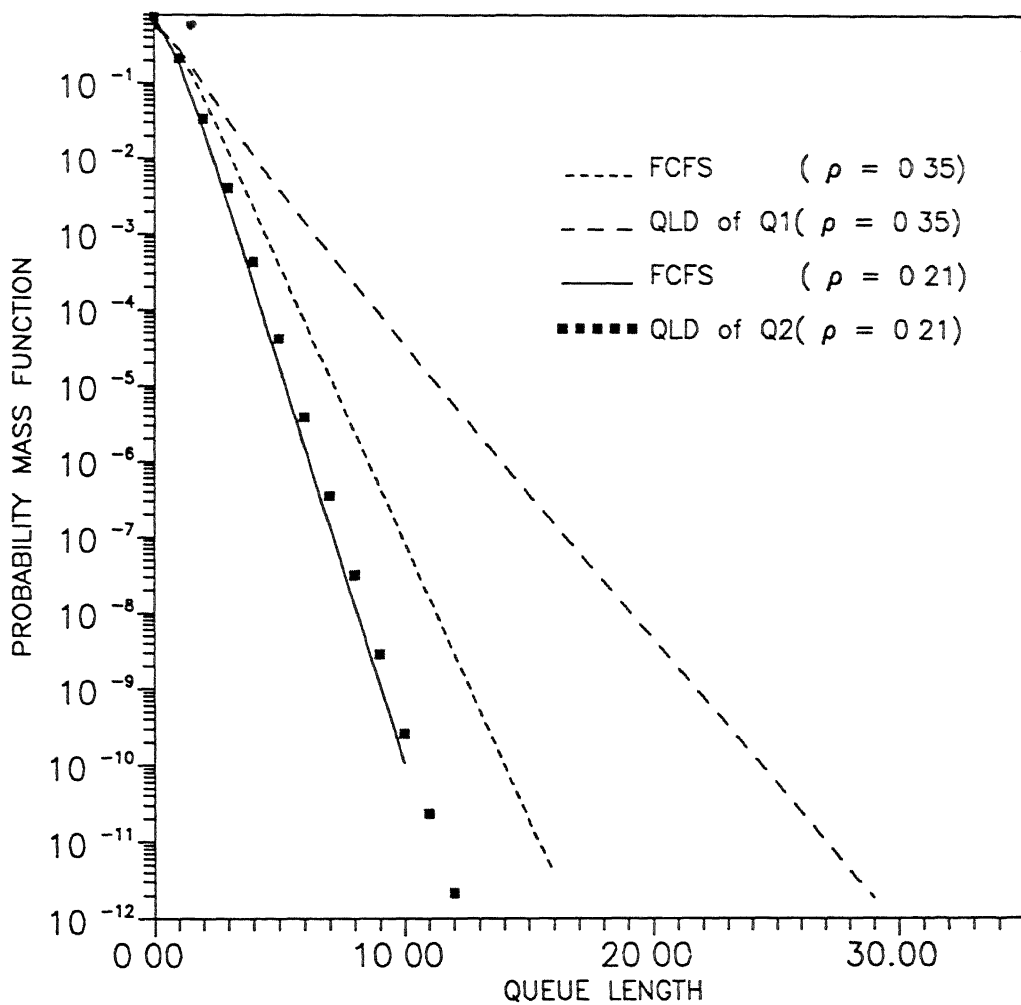


FIG 5.46 COMPARISON OF THE QLDs OF Q1 AND Q2 WITH FCFS QUEUES WITH DEDICATED SERVERS FOR TRAFFICS FROM (150,90) TYPE1 SOURCES

CHAPTER 6

QUEUEING DELAY OF A NON-PREEMPTIVE MMPP/D/1 DUAL PRIORITY SYSTEM

6.1. INTRODUCTION

In this chapter, the evaluation of the Laplace Steiltjes Transform (LST) of the virtual waiting time of the cells arriving at Q2, the higher priority queue and the computation of the average delays at Q1 and Q2 are considered. The average delays at Q1 and Q2 are evaluated for some typical examples and the results are compared with those obtained using simulations. The validation of these results are also carried out by extending these results for the special case of Poisson arrivals at both Q1 and Q2.

Notations used in this chapter are the same as those used in the previous chapters. For example, $X''(t)$, $J''(t)$ denote the number of cells in Q2 and the phase of MMPP 2 at time t . Similarly, $X'(t)$, $J'(t)$ denote the number of cells in Q1 and the phase of MMPP 1 at time t . Given the model parameters of MMPP 1 and MMPP 2, the parameters of MMPP 1 and MMPP 2 are again given by equations (3.2.5) - (3.2.6). Dual priority classes and non-preemptive priority discipline are also assumed. In addition to Q1 and Q2, we shall consider a third queue denoted as Q. The traffic arriving at both Q1 and Q2 are assumed to be fed to Q and are served by a single server on a FCFS basis. The traffic to Q is modelled as a MN phase MMPP. It may be recalled that the total number of phases of MMPP 1, MMPP 2 are M , N respectively. The number of cells in Q system and the phase of the composite MMPP at time t is denoted as $X(t)$ and $J(t)$ respectively.

6.2. VIRTUAL WAITING TIME DISTRIBUTION OF A HIGH PRIORITY CELL

The knowledge of the distribution of the virtual waiting time of a customer in a queueing system is desirable for several reasons. Firstly, it enables one to find the probability that the queueing delay exceeds a certain threshold. The cells whose delays have exceeded this threshold may be dropped as they exceed either the maximum delay or the maximum delay jitter limit and hence will not be of use at the destination. The cell loss probability arising out of these cell discards may have to be kept under control. Secondly, knowing the distribution, the average delays can be computed. This in turn can be used to find the average queue length using Little's formula. However, it should be mentioned here that the computation of these statistics directly using the expression for the distribution of the virtual waiting time is often difficult and time consuming. Hence, using this expression the LST of the distribution function is found first and then the computation of the required statistics from this transform becomes simpler and more manageable. In this section we consider the computation of the virtual waiting time of a customer arriving at Q2, i.e. the higher priority queue.

The virtual waiting time of a cell arriving at Q2 at time t is the time it waits before entering service. Let this be denoted as $v''(t)$. It may be noted that the virtual waiting time for a cell arriving at time t is non-zero if either a Q1 cell or a Q2 cell is undergoing service at t . The duration of $v''(t)$ depends in turn on the number of cells in Q1 and Q2 at time t .

In a manner similar to that in Ramaswami[1], we define an $MN \times 1$ vector $W''(\sigma)$, to characterize the cdf of $v''(t)$. The ℓ^{th} element of $W''(\sigma)$ is denoted as $W''_{\ell}(\sigma)$ and it gives the probability that $v''(t) \leq \sigma$ and MMPP $\underline{2}$ is in phase ℓ . For $\sigma \geq 0$ and $1 \leq \ell \leq MN$, $W''_{\ell}(\sigma)$ is obtained as the limit of the conditional probability given by-

$$W_{\ell}''(\sigma) = \lim_{t \rightarrow \infty} P[v''(t) \leq \sigma, \underline{j}(t) = \ell | X''(0) = i, \underline{j}(0), X'(0) = i'] \quad (6.2.1)$$

It may be noted that $W_{\ell}''(0)$ gives the corresponding conditional probability when the virtual waiting time is zero. Then (6.2.1) can be rewritten by separating it for the case where the virtual waiting time is zero and the case where it is non-zero. This yields -

$$W_{\ell}''(\sigma) = W_{\ell}''(0) + \lim_{t \rightarrow \infty} P[0 < v''(t) \leq \sigma, \underline{j}(t) = \ell | X''(0) = i, \underline{j}(0), X'(0) = i'] \quad (6.2.2)$$

For ease of notation we define $W_{\ell}''(\sigma, t)$ as follows -

$$W_{\ell}''(\sigma, t) = P[0 < v''(t) \leq \sigma, \underline{j}(t) = \ell | X''(0) = i, \underline{j}(0), X'(0) = i'] \quad (6.2.3)$$

The conditional probability given in (6.2.3) can be evaluated by considering all the possible cases under which $v''(t)$ is non-zero. Let τ be the time at which the latest departure from either Q1 or Q2 occurred before time t . Then $v''(t)$ is non-zero under the following cases

Case 1 a Q2 cell departed at τ and $\{X''(\tau) > 0 \text{ and } X'(\tau) \geq 0\}$

Case 2 a Q2 cell departed at τ and $\{X''(\tau) = 0 \text{ and } X'(\tau) > 0\}$

Case 3 a Q1 cell departed at τ and $\{X''(\tau) > 0 \text{ and } X'(\tau) \geq 0\}$

Case 4 a Q1 cell departed at τ and $\{X''(\tau) = 0 \text{ and } X'(\tau) > 0\}$

Case 5 either a Q2 or a Q1 cell departed at τ , $\{X''(\tau) = X'(\tau) = 0\}$ and at least one cell arrives either at Q1 or at Q2 in (τ, t)

The conditional probability corresponding to each of the five cases referred to above can be computed by considering the following chain of conditional events E1, E2 and E3 defined as follows

E1 $\{X''(\tau) = c_1, X'(\tau) = c_1' \text{ and } \underline{j}(\tau) = j | X''(0) = i, X'(0) = i' \text{ and } \underline{j}(0) = j'\}$

E2 $\{X''(t) = c_1 + c_2 \text{ and } \underline{j}(t) = \ell | X''(\tau) = c_1, X'(\tau) = c_1' \text{ and } \underline{j}(\tau) = j\}$

E3 $\{v''(t) \leq \sigma | X''(t) = c_1 + c_2 \text{ and } \underline{j}(t) = \ell\}$

where i, i', c_1, c_1', c_2 are integers greater than or equal to zero. Their actual value depends on which of the five cases is true. The steps involved in computing the conditional probability corresponding to each of these cases is

lengthy. Hence we consider the steps corresponding to case 1 here and relegate the details corresponding to the other cases to Appendix (6 A)

To evaluate $W_l''(\sigma, t)$ under the condition that case 1 is true, we compute the probability of occurrence of the events E1, E2 and E3 corresponding to case 1 as follows. First let us compute the probability of occurrence of event 1 denoted as $P[E1]$. In this case, as a Q2 cell departs at τ and leaves Q2 non empty $c_1 > 0$ and $c_1' \geq 0$. The condition $c_1' \geq 0$ implies that $X'(\tau) \geq 0$. This in turn implies that $P[X'(\tau)=c_1' | X'(0)=1'] = 1$ and $P(E1)$ is independent of $X'(\tau)$. Hence $P[E1]$ can be computed by considering the event $\{X''(\tau) = c_1 (c_1 > 0), j(\tau) = j | X''(0)=1, j(0)=j'\}$. For the Markov Renewal Process $Q''()$, the probability of this event is also equal to the expected number of visits of $Q''()$ to the state (c_1, j) at time τ , given that at time 0 the state was $(1, j')$. Hence $P[E1]$ is given by

$$P(E1) \Big|_{\text{Case 1}} = d\phi_{c_1 j}^{1j'}(\tau) \quad (6.2.4)$$

where

$$\phi_{c_1 j}^{1j'}(\tau) = E[\text{No of visits of } Q''() \text{ to the state } (c_1, j) \text{ in } [0, \tau] | \text{ at time 0 the state was } (1, j')]$$

Next, the event 2 is considered. In case 1, $X''(t)$ and $j(t)$ do not depend on the value of $X'(\tau)$. In view of this $P(E2)$ can be evaluated as follows

$$\begin{aligned} P(E2) \Big|_{\text{Case 1}} &= P[X''(t)=c_1+c_2, j(t)=\ell | X''(\tau) = c_1, j(\tau) = j \text{ and } X'(\tau) \geq 0] \\ &= P[X''(t)=c_1+c_2, j(t)=\ell | X''(\tau) = c_1, j(\tau) = j] \\ &= P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) \end{aligned} \quad (6.2.5)$$

Since the service time/cell is D sec, $(t-\tau)$, the duration of service received by the cell under service, can utmost be D and the factor $u(D-t+\tau)$ ensures that $P[E2]$ is zero for $(t-\tau) \geq D$. The matrix $P''(c_2, t)$, whose (j, ℓ) th element is $P_{j\ell}''(c_2, t)$, is defined in Sec 3.3

Next, the evaluation of $P[E3]$ is considered. The virtual waiting time $v''(t)$ at Q2 in this case is the sum of w , the residual service time (RST) of the on going service and the service time required for the remaining c_1+c_2-1 cells. Let $H(\sigma)$, $H^{<n>}(\sigma)$ denote the cdf of the service time and the n fold convolution of $H(\sigma)$ with itself, respectively. It may be noted that for Q2, the cdf of the interdeparture time of cells from a non-empty Q2 and the service time cdf are equal. Due to the constant service time/cell, $dH(t)$ is a delta function and hence we get

$$P[E3] = P[v''(t) \leq \sigma \mid X''(t)=c_1+c_2, \underline{j}(t)=\underline{\ell}] \quad (6.2.6)$$

$$= \int_{w=0}^{\sigma} dH(t-\tau+w) H^{<c_1+c_2-1>}(\sigma-w) \quad (6.2.7)$$

$$P(E3) \Big|_{\text{Case 1}} = u(\sigma+t-\tau-c_1D-c_2D) \quad (6.2.8)$$

The product of $P[E1]$, $P[E2]$ and $P[E3]$ gives $W''_{\ell}(\sigma, t)$ under the conditions that case 1 is true, previous departure occurs at time τ , MMPP $\underline{2}$ is in phase j at τ , number of customers in Q2 at τ is c_1 and the number of arrivals in (τ, t) is c_2 . The condition on the departure instant can be removed by integrating w.r.t τ . The other random variables are discrete random variables and hence the conditions on them can be removed by performing the summation over their admissible range in case 1. Removing these conditions we get

$$W''_{\ell}(\sigma, t) \Big|_{\text{case 1}} = \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \int_{\tau=0}^t d\phi_{c_1j}^{1j'}(\tau) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1D-c_2D) \quad (6.2.9)$$

As mentioned earlier, $W''_{\ell}(\sigma, t)$ is computed corresponding to the cases 2 to 5 in Appendix (6.A). Summing up the expressions corresponding to all the five cases, $W''_{\ell}(\sigma, t)$ is given by

$$W''_{\ell}(\sigma, t) = P[0 < v''(t) < \sigma, \underline{j}(t)=\underline{\ell} \mid X''(0)=1, \underline{j}(0)=j', X'(0)=1']$$

$$\begin{aligned}
&= \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \int_0^t d\phi_{c_1j}^{1j'}(\tau) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1D-c_2D) \\
&+ (1-P_0') \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \int_0^t d\phi_{0j}^{1j'}(\tau) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-D-c_2D) \\
&+ (1-P_0') \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{j'=1}^{MN} \sum_{c_2=0}^{\infty} \int_{\tau'=0}^{t-D} \left\{ \int_{\tau=\tau'+D}^t d\phi_{0j'}^{1j'}(\tau') dU_{c_1j}''(\tau-\tau') \right\} \\
&\quad P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1D-c_2D) \Big] \\
&+ \sum_{c_1=1}^{\infty} \sum_{k=1}^{MN} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \int_0^t d\phi_{c_1k}^{1'k'}(\tau) y''(0, j) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-D-c_2D) \\
&+ \sum_{j'=1}^{MN} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \int_0^t d\phi_{0j'}^{1j'}(\tau) \int_{t'=0}^{t-\tau} dU_{1j}''(t-\tau-t') P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-c_2D-D)
\end{aligned}$$

(6.2.10)

where

$[P''(n, t)]_{1j}$	$P[n \text{ cells arrive at } Q2 \text{ in } (0, t], j(t)=j \mid j(0)=1]$
$[U_k''(t)]_{1j}$	$P[\text{BP of } Q2 \text{ starts at or before } t, Q1 \text{ non-empty when the 1st cell arrives at } Q2, X''(t)=k, j(t)=j \mid X''(0)=0, j(0)=1]$
$[U_1(t)]_{1j}$	$P[\text{BP of } Q \text{ starts at or before } t, j(t)=j \mid X(0)=0, j(0)=1]$
$u(t)$	unit step function
$\phi_{c_1j}^{1j'}(\tau)$	$E[\text{No of visits of } Q'(\cdot) \text{ to the state } (c_1, j) \text{ in } [0, \tau] \mid \text{at time } 0 \text{ the state was } (1, j')]$
$\phi_{c_1j}^{1j'}(\tau)$	$E[\text{No of visits of } Q(\cdot) \text{ to the state } (c_1, j) \text{ in } [0, \tau] \mid \text{at time } 0 \text{ the state was } (1, j')]$
$y''(0, j)$	$P[X''(t) = 0, j(t)=j] \text{ at an arbitrary time } t$

P'_0 $P[X'(t) = 0]$ at an arbitrary time t

Next, $W''_\ell(\sigma)$ is obtained by evaluating (6.2.10) as $t \rightarrow \infty$ and adding the term corresponding to the case when $v''(t)$ is zero. The latter term is obviously equal to $W''_\ell(0)$. The limit of (6.2.10) is evaluated by applying the key renewal theorem (KRT) to each of the terms of (6.2.10). To save the details in this step, we go through the steps in applying the KRT to the first term of (6.2.10) here and give the steps required for the other terms of (6.2.10) in Appendix (6.B).

The key renewal theorem (see for e.g. Wolff [2]) states that

if $R(t)$ is any directly Riemann integrable function and $\phi_\iota(t)$ is the renewal function of the state ι of a renewal process (i.e. average number of renewals of the state ι in $(0, t]$) then

$$\lim_{t \rightarrow \infty} \int_0^t d\phi_\iota(\tau) R(t-\tau) = \frac{1}{m_\iota} \int_0^\infty R(t) dt \quad (6.2.11)$$

where m_ι is the mean recurrence time of the state ι of the renewal process.

Generalizing this result to the two dimensional Markov renewal process $Q''(\cdot)$ and applying the KRT to the first term of (6.2.10) we get

$$\begin{aligned} \lim_{t \rightarrow \infty} \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \int_0^t \left[d\phi_{c_1 j}''(\tau) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1 D-c_2 D) \right] \\ = \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m''(c_1, j)} \int_0^\infty P_{j\ell}''(c_2, t) u(D-t) u(\sigma+t-c_1 D-c_2 D) \quad (6.2.12) \end{aligned}$$

where $m''(c_1, j)$ is the mean recurrence time (MRT) of the state (c_1, j) of $Q''(\cdot)$ and t is a dummy variable. Let the mean recurrence time of the state (c, k) of $Q(\cdot)$ and $Q'(\cdot)$ be denoted as $m(c, k)$ and $m'(c, k)$ respectively. Applying the KRT to all the terms of (6.2.10), using (6.2.12) and Appendix (6.B) we get

$$\begin{aligned}
 W_{\ell}''(\sigma) = & W_{\ell}''(0) + \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m''(c_1, j)} \int_{t=0}^{\infty} P_{j\ell}''(c_2, t) u(D-t) u(\sigma+t-c_1D-c_2D) \\
 & + (1-p_0') \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m''(0, j)} \int_{t=0}^{\infty} P_{j\ell}''(c_2, t) u(D-t) u(\sigma+t-D-c_2D) \\
 & + (1-p_0') \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{j'=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m''(0, j')} \int_{v'=0}^D U_{c_1}''(\infty) P_{j\ell}''(c_2, v') u(\sigma+v'-c_1D-c_2D) \\
 & - (1-p_0') \sum_{c_1=1}^{\infty} \sum_{j=1}^{MN} \sum_{j'=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m''(0, j')} \int_{t=0}^{\infty} \int_{v=0}^D dU_{c_1}''(v) P_{j\ell}''(c_2, t-v) \\
 & \quad u(D-t+v) u(\sigma+t-v-c_1D-c_2D) \\
 & + \sum_{c_1=1}^{\infty} \sum_{k=1}^{MN} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m'(c_1, k)} \int_{t=0}^{\infty} y''(0, j) P_{j\ell}''(c_2, t) u(D-t) u(\sigma+t-D-c_2D) \\
 & + \sum_{j'=1}^{MN} \sum_{j=1}^{MN} \sum_{c_2=0}^{\infty} \frac{1}{m(0, j')} \int_{t'=0}^D U_1(\infty) P_{j\ell}''(c_2, t') u(\sigma+t'-D-c_2D) \quad (6.2.13)
 \end{aligned}$$

Next the m r t of the states of $Q''()$, $Q'()$ and $Q()$ appearing in (6.2.13) is eliminated as follows. Using (5.5.25) and (5.5.34) it can be verified that the m r t of the state (c, j) of the Markov renewal processes $Q'()$, $Q''()$ and $Q()$ can be expressed in terms of the stationary probabilities of the state (c, j) of the Markov chains $Q'(\infty)$, $Q''(\infty)$ and $Q(\infty)$ as follows -

$$m'(c, j) = [\xi^{*'} x'(c, j)]^{-1} \quad (6.2.14)$$

$$m''(c, j) = [\xi^{*''} x''(c, j)]^{-1} \quad (6.2.15)$$

$$m(c, j) = [\xi^* x(c, j)]^{-1} \quad (6.2.16)$$

where $\xi^{*''}$, $\xi^{*'}$ and ξ^* are the inverse of the mean sojourn times of the

semi-Markov processes $Q''()$, $Q'()$ and $Q()$ respectively. It may be noted that (6.2.16) can be obtained from (5.5.25) by replacing the parameters of Q_2 by those of Q and substituting p'_0 to be 0. In the first and the second terms of (6.2.13), the upper limit on t can be changed to be D as

$$u(D-t)=0 \quad \text{for } t \geq D \quad (6.2.17)$$

Let the $1 \times MN$ vector, whose l th element is $W_l''(\sigma)$, be denoted as $W''(\sigma)$. Using (6.2.14)-(6.2.17) in (6.2.13) and writing it in matrix form we get

$$\begin{aligned} W''(\sigma) = & \sum_{c_1=1}^{\infty} \xi^{*''} \sum_{c_2=0}^{\infty} x_{c_1}'' \int_0^D P''(c_2, t) u(\sigma+t-c_1 D-c_2 D) \\ & + (1-p'_0) \sum_{c_2=0}^{\infty} \xi^{*''} x_0'' \int_0^D P''(c_2, t) u(\sigma+t-D-c_2 D) \\ & + (1-p'_0) \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*''} x_0'' \int_0^D \tilde{U}_{c_1}'' P''(c_2, t) u(\sigma+t-c_1 D-c_2 D) \\ & - (1-p'_0) \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*''} x_0'' \int_0^{\infty} \int_0^D dU_{c_1}''(v) P''(c_2, t-v) u(D-t+v) u(\sigma+t-v-c_1 D-c_2 D) \\ & + \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*'} x_{c_1}' e^{-y_0''} \int_0^D P''(c_2, t) u(\sigma+t-D-c_2 D) \\ & + \sum_{c_2=0}^{\infty} \int_0^D \xi^* x_0 \tilde{U}_1 P''(c_2, t) u(\sigma+t-D-c_2 D) + W''(0) \end{aligned} \quad (6.2.18)$$

Here, $W''(0)$ denotes the probability that $v''(t)$ is zero as $t \rightarrow \infty$ and $\tilde{U}_{c_1}'', \tilde{U}_1$ denote the LST of $U_{c_1}''(t)$ and $U_1(t)$ respectively evaluated at $s=0$. It may be noted that they are also equal to $U_{c_1}''(\infty)$ and $U_1(\infty)$ respectively.

6.2. LST OF THE VIRTUAL WAITING TIME DISTRIBUTION

In this section the Laplace Steiltjes Transform of $W''(\sigma)$, the c.d.f. of the virtual waiting time $v''(t)$ is obtained. As noted earlier, this will be useful for the computation of the percentile of the queueing delays at Q_2 as

well as the first moments of the queueing delays at Q1 and Q2. We denote the LST of $W''(\sigma)$ as $\tilde{W}''(s)$

We compute the Laplace transform of the various terms of (6.2.18) first. The LST of $W''(\sigma)$ is then computed using the relation that the LST of a distribution function is equal to s times the Laplace transform of the distribution function, i.e.

$$\int_0^{\infty} dW''(\sigma) e^{-s\sigma} = s \int_0^{\infty} W''(\sigma) d\sigma e^{-s\sigma} \quad (6.3.1)$$

We start with the computation of the Laplace Transform (LT) of the 1st term of (6.2.18). Let the LT of the i th term on the RHS of (6.2.18) be denoted as T_i . T_1 can be simplified as follows

$$T_1 = \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*n} x_{c_1}'' \int_{t=0}^D P''(c_2, t) \int_{\sigma=0}^{\infty} u(\sigma + t - c_1 D - c_2 D) e^{-s\sigma} d\sigma \quad (6.3.2)$$

$$= \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*n} x_{c_1}'' \int_{t=0}^D P''(c_2, t) \frac{1}{s} e^{-s[c_1 D + c_2 D - t]} dt \quad (6.3.3)$$

$$= \sum_{c_1=1}^{\infty} \xi^{*n} x_{c_1}'' [e^{-sD}]^{-c_1} \int_{t=0}^D \left[\sum_{c_2=0}^{\infty} P''(c_2, t) [e^{-sD}]^{-c_2} \right] \frac{1}{s} e^{st} dt \quad (6.3.4)$$

Let the z transform of x_n'' evaluated at $z = e^{-sD}$ be denoted as $\tilde{x}''(e^{-sD})$. Using (3.6.1) in (6.3.4) we get

$$T_1 = \frac{\xi^{*n}}{r} [\tilde{x}''(r) - x_0''] \int_{t=0}^D \frac{1}{s} dt e^{[R''(r) + sI]t} \quad (6.3.5)$$

Here, $r = e^{-sD}$, I is the $MN \times MN$ identity matrix and $R''(z)$ is given by (3.6.1)

Let the z transform of the LST of $U_k''(t)$ and $A_k''(t)$ evaluated at $z=r$ and $s=0$ be denoted as $\tilde{U}''(r, 0)$ and $\tilde{A}''(r, 0)$ respectively. For simplification of (6.3.5) we use the expression for $\tilde{A}''(r, 0)$ given by (5.4.5) which is reproduced here for

convenience

$$\mathcal{X}''(z) = x_0'' [\tilde{U}''(z, 0) - I] \tilde{A}''(z, 0) [zI - \tilde{A}''(z, 0)]^{-1} \quad (6.3.6)$$

Using (6.3.6) in (6.3.5) and integrating w.r.t. t we get

$$T1 = \xi^{*''} [\mathcal{X}''(r) - x_0''] \frac{1}{s} \left[e^{[R''(r) + sI]D} - I \right] [R''(r) + sI]^{-1} \quad (6.3.7)$$

$$= \xi^{*''} [\mathcal{X}''(r) - x_0''] \frac{1}{rs} [\tilde{A}''(r, 0) - rI] [R''(r) + sI]^{-1} \quad (6.3.8)$$

$$= \xi^{*''} x_0'' \left[[\tilde{U}''(r, 0) - I] \tilde{A}''(r, 0) [rI - \tilde{A}''(r, 0)]^{-1} - I \right]$$

$$\frac{1}{rs} [\tilde{A}''(r, 0) - rI] [R''(r) + sI]^{-1} \quad (6.3.9)$$

$$= \xi^{*''} x_0'' \left\{ [\tilde{U}''(r, 0) - I] \tilde{A}''(r, 0) [rI - \tilde{A}''(r, 0)]^{-1} [\tilde{A}''(r, 0) - rI] \right.$$

$$\left. - I [\tilde{A}''(r, 0) - rI] \right\} \frac{1}{rs} [R''(r) + sI]^{-1} \quad (6.3.10)$$

$$= \xi^{*''} x_0'' \left\{ -[\tilde{U}''(r, 0) - I] \tilde{A}''(r, 0) - I [\tilde{A}''(r, 0) - rI] \right\} \frac{1}{rs} [R''(r) + sI]^{-1} \quad (6.3.11)$$

$$= \left[\xi^{*''} x_0'' [I - \tilde{U}''(r, 0) \tilde{A}''(r, 0) e^{sD}] \right] \frac{1}{s} [R''(r) + sI]^{-1} \quad (6.3.12)$$

This completes the simplification of T1

In order to avoid getting lost into the details of the evaluation of the Laplace Transform of the remaining terms of (6.2.18), we would like mention here that the steps are similar and one may refer to (6.3.41) for the final expression

Next, the simplification of T2 is carried out

$$T2 = (1-p'_0) \sum_{c2=0}^{\infty} \xi^{*n} x_0'' \int_{t=0}^D P''(c2, t) \int_{\sigma=0}^{\infty} u(\sigma+t-D-c2D) e^{-s\sigma} d\sigma \quad (6.3.13)$$

As $u(\cdot)$ is zero for $\sigma < D + c2D - t$ the upper limit on σ can be changed to be $D + c2D - t$. Integrating (6.3.13) w.r.t. σ we get

$$T2 = (1-p'_0) \sum_{c2=0}^{\infty} \xi^{*n} x_0'' \int_{t=0}^D P''(c2, t) \frac{1}{s} e^{-s[D+c2D-t]} dt \quad (6.3.14)$$

Let $r = e^{-sD}$. Combining the like terms and using (3.6.1) we get

$$T2 = (1-p'_0) \xi^{*n} x_0'' r \int_{t=0}^D \left[\sum_{c2=0}^{\infty} P''(c2, t) r^{c2} \right] \frac{1}{s} e^{st} dt \quad (6.3.15)$$

$$= (1-p'_0) \xi^{*n} x_0'' r \int_{t=0}^D e^{[R''(r)+sI]t} \frac{1}{s} dt \quad (6.3.16)$$

$$= (1-p'_0) \xi^{*n} x_0'' \left[\tilde{A}''(r, 0) - rI \right] \frac{1}{s} \left[R''(r) + sI \right]^{-1} \quad (6.3.17)$$

$T3$, the LT of the 3rd term of (6.2.18) can be obtained by proceeding along the same lines as for $T1$ and the steps are as follows

$$T3 = (1-p'_0) \xi^{*n} x_0'' \sum_{c1=1}^{\infty} \sum_{c2=0}^{\infty} U''_{c1} \int_{t=0}^D P''(c2, t) \int_{\sigma=0}^{\infty} u(\sigma+t-c1D-c2D) e^{-s\sigma} d\sigma \quad (6.3.18)$$

$$= (1-p'_0) \xi^{*n} x_0'' \sum_{c1=1}^{\infty} \sum_{c2=0}^{\infty} U''_{c1} \int_{t=0}^D P''(c2, t) \frac{1}{s} e^{-s[c1D+c2D-t]} dt \quad (6.3.19)$$

$$= \xi^{*n} x_0'' \left\{ (1-p'_0) \sum_{c1=1}^{\infty} U''_{c1} r^{-c1} \right\} \int_{t=0}^D \left[\sum_{c2=0}^{\infty} P''(c2, t) r^{-c2} \right] \frac{1}{s} e^{st} dt \quad (6.3.20)$$

Using (6.2.11) U''_{c1} is given by-

$$U''_{c1} = U''_{c1}(\infty) \Big|_{Q1 \text{ not empty}} = \tilde{U}''_{c1}(s) \Big|_{Q1 \text{ not empty}, s=0} \quad (6.3.21)$$

Using (6.3.21), (3.6.1) and (3.6.18)-(3.6.20) we get-

$$(1-p'_0) \sum_{c_1=1}^{\infty} U''_{c_1} r^{-c_1} = \tilde{U}''(r,0) - p'_0 [\Lambda'' - Q^*]^{-1} \Lambda'' r \quad (6.3.22)$$

Using (6.3.22) and (3.6.1) in (6.3.20) and integrating w.r.t. t we get

$$T_3 = (1-p'_0) \xi^{**} x''_0 \left\{ U''(r,0) - p'_0 [\Lambda'' - Q^*]^{-1} \Lambda'' r \right\} \int_{t=0}^D \frac{1}{s} dt e^{[R''(r)+sI]t} \quad (6.3.23)$$

$$= \xi^{**} x''_0 \left\{ U''(r,0) - p'_0 [\Lambda'' - Q^*]^{-1} \Lambda'' r \right\} \frac{1}{rs} [\tilde{\Lambda}''(r,0) - rI] [R''(r)+sI]^{-1} \quad (6.3.24)$$

Using (5.5.27) in (6.3.24) and simplifying we get

$$T_3 = \left\{ \xi^{**} x''_0 [\tilde{U}''(r,0) \tilde{\Lambda}''(r,0) e^{sD} - \tilde{U}''(r,0)] - y''_0 \Lambda'' p'_0 \right\} \frac{1}{s} [R''(r)+sI]^{-1} \quad (6.3.25)$$

Next, T_4 , the LT of the 4th term of (6.2.18) is obtained as follows

$$T_4 = (1-p'_0) \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{**} x''_0 \int_{t=0}^{\infty} \int_{v=0}^D dU''_{c_1}(v) P''(c_2, t-v) u(D-t+v) \int_{\sigma=0}^{\infty} u(\sigma+t-v-c_1 D-c_2 D) e^{-s\sigma} d\sigma \quad (6.3.26)$$

Integrating (6.3.26) w.r.t. σ we get-

$$= (1-p'_0) \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{**} x''_0 \int_{t=0}^{\infty} \int_{v=0}^D dU''_{c_1}(v) P''(c_2, t-v) u(D-t+v) \frac{1}{s} e^{-s[c_1 D+c_2 D-t+v]} dt \quad (6.3.27)$$

Carrying out the summation w.r.t. c_2 and using (3.6.1) we get-

$$= (1-p'_0) \xi^{**} x''_0 \sum_{c_1=1}^{\infty} \int_{t=0}^{\infty} \int_{v=0}^D dU''_{c_1}(v) r^{c_1} \sum_{c_2=0}^{\infty} P''(c_2, t-v) r^{c_2} u(D-t+v) \frac{1}{s} e^{-s[-t+v]} dt \quad (6.3.28)$$

$$= (1-p'_0) \xi^{**} x''_0 \sum_{c_1=1}^{\infty} \int_{t=0}^{\infty} \int_{v=0}^D dU''_{c_1}(v) r^{c_1} u(D-t+v) \frac{1}{s} dt e^{[R''(r)+sI](t-v)} \quad (6.3.29)$$

Note that (6.3.29) is separable in t and v and hence the order of integration w.r.t. v and t can be changed. Because of the term $u(D-t+v)$ the upper limit of t can be changed to be $D+v$. Integrating w.r.t. t we obtain-

$$= (1-p'_0)\xi^{**}x_0'' \sum_{c_1=1}^{\infty} \int_{v=0}^D dU_{c_1}''(v)r^{c_1} e^{-v[R''(r)+sI]} \int_{t=0}^{D+v} dt e^{[R''(r)+sI]t} \quad (6.3.30)$$

$$= (1-p'_0)\xi^{**}x_0'' \sum_{c_1=1}^{\infty} \int_{v=0}^D dU_{c_1}''(v)r^{c_1} e^{-v[R''(r)+sI]} \left\{ e^{[R''(r)+sI](D+v)} - 1 \right\} \frac{1}{s} [R''(r)+sI]^{-1} \quad (6.3.31)$$

Using (6.2.11), (3.4.6) and (3.6.1) the summation wrt c_1 in (6.3.31) can be simplified. Noting that the maximum value of v is D , we get

$$\sum_{c_1=1}^{\infty} dU_{c_1}''(v)r^{c_1} = \frac{1}{D} \int_{w=0}^v P''(0, v-w) \Lambda'' dw \sum_{c_1=1}^{\infty} P''(c_1-1, w)r^{c_1} \quad (6.3.32)$$

$$= \frac{1}{D} \int_{w=0}^v P''(0, v-w) \Lambda'' dw r e^{[R''(r)]w} \quad (6.3.33)$$

Substituting (6.3.33) in (6.3.31) we get

$$T4 = (1-p'_0)\xi^{**}x_0'' \int_{v=0}^D \frac{1}{D} \int_{w=0}^v P''(0, v-w) \Lambda'' dw r e^{(w-v)R''(r)} e^{-vs} \left\{ e^{[R''(r)+sI](D+v)} - 1 \right\} \frac{1}{s} [R''(r)+sI]^{-1} \quad (6.3.34)$$

T4 has to be evaluated using (6.3.34) by numerical integration. It may be noted that the maximum value of v is D . When D is small (it is of the order of few microsecond for a cell size of 53 bytes and output link capacity of 150 Mbps) and the traffic offered at Q2 is not too large, the terms inside the first bracket of (6.3.34) become almost equal and hence T4 may be neglected as a good approximation.

The computation of T5, the LT of the 5th term of (6.2.18) can be carried out along the same lines as that of T2 and is given by

ξ^{**} , ξ^{*} and ξ^* can be eliminated using the following relations obtained in Sec 5.5

$$y_0'' = \xi^{**} x_0'' [\underline{\Lambda}'' - \underline{Q}^*]^{-1} \quad (6.3.42)$$

$$y_0' = \xi^{*} x_0' [\underline{\Lambda}' - \underline{Q}^*]^{-1} \quad (6.3.43)$$

$$y_0 = \xi^* x_0 [\underline{\Lambda} - \underline{Q}^*]^{-1} \quad (6.3.44)$$

Using these equations, (6.3.41) can be written as

$$\tilde{W}''(s) = W''(0) + \left[y_0'' [\underline{\Lambda}'' - \underline{Q}^*] [-\tilde{U}''(r, 0)] + W_c [\tilde{A}''(r, 0) - rI] + \mathcal{E}(s) \right] \left[R''(r) + sI \right]^{-1} \quad (6.3.45)$$

$$W_c = (1 - p_0'') y_0'' (\underline{\Lambda}'' - \underline{Q}^*) + \frac{1}{x_0' e} y_0' (\underline{\Lambda}' - \underline{Q}^*) e (1 - x_0' e) y_0'' + y_0 \underline{\Lambda} - y_0'' \underline{\Lambda} p_0' \quad (6.3.46)$$

$$\mathcal{E}(s) = (1 - p_0') y_0'' (\underline{\Lambda}'' - \underline{Q}^*) \int_{v=0}^D \frac{1}{D} \int_{w=0}^v P''(0, v-w) \underline{\Lambda}'' dw e^{(w-v)R''(r)} e^{-vs} \left\{ e^{[R''(r) + sI](D+v)} - I \right\} \quad (6.3.47)$$

6.4. VIRTUAL WAITING TIME DISTRIBUTION AND ITS LST USING THE APPROXIMATE MODEL

We shall use the same symbols to denote the parameters of the approximate model as those used for the corresponding parameters in the exact model. As in the exact model, we define an $MN \times 1$ vector $W''(\sigma)$, to characterize the cdf of $v''(t)$

The ℓ^{th} element of $W''(\sigma)$ is denoted as $W_\ell''(\sigma)$ and it gives the probability that $v''(t) \leq \sigma$ and MMPP 2 is in phase ℓ . For $\sigma \geq 0$ and $1 \leq \ell \leq MN$, $W_\ell''(\sigma)$ is obtained as the limit of the conditional probability given by-

$$W_\ell''(\sigma) = \lim_{t \rightarrow \infty} P[v''(t) \leq \sigma, J''(t) = \ell | X''(0) = i, J''(0) = j, X'(0) = i', J'(0)] \quad (6.4.1)$$

where $J''(t)$ and $J'(t)$ denote the phases of the MMPPs to Q2 and Q1 at time t , respectively

It may be noted that $W_\ell''(0)$ gives the corresponding conditional

probability when the virtual waiting time is zero Then (6.2.1) can be rewritten by separating the case where the virtual waiting time is zero as follows

$$W_{\ell}''(\sigma) = \lim_{t \rightarrow \infty} P[0 < v''(t) \leq \sigma, J''(t) = \ell | \lambda''(0) = 1, J''(0) = j'', X'(0) = i', J'(0) = j'] + W_{\ell}''(0) \quad (6.4.2)$$

For ease of notation we define $W_{\ell}''(\sigma, t)$ as follows -

$$W_{\ell}''(\sigma, t) = P[0 < v''(t) \leq \sigma, J''(t) = \ell | \lambda''(0) = 1, J''(0) = j'', X'(0) = i', J'(0) = j'] \quad (6.4.3)$$

The conditional probability given by (6.4.3) has to be computed corresponding to each of the five cases under which $v''(t)$ is non-zero (These five cases are listed in Sec 6.2) This can be achieved by computing the probability of occurrence of the following chain of conditional events E1, E2 and E3 defined as -

$$E1 \{ X''(\tau) = c_1, X'(\tau) = c_1', J''(\tau) = j | \lambda''(0) = 1, J''(0) = j'', X'(0) = i', J'(0) = j' \}$$

$$E2 \{ X''(t) = c_1 + c_2 \text{ and } J''(t) = \ell | X''(\tau) = c_1, X'(\tau) = c_1' \text{ and } J''(\tau) = j \}$$

$$E3 \{ v''(t) \leq \sigma | X''(t) = c_1 + c_2 \text{ and } J''(t) = \ell \}$$

where 1, 1', c1, c1', c2 are integers greater than or equal to zero Their actual values depend on which of the five cases is true Let us first compare the events E2, E3 corresponding to the approximate model with those of exact model It can be seen that these events differ only in the phase terms viz $\underline{j}(\)$ is replaced by $J''(\)$ in the present case Hence the computation of $P(E2)$ and $P(E3)$ can be achieved by proceeding along the same lines as for the exact model In each case $\underline{j}(\)$ is replaced by $J''(\)$ However, in the 5th case $P(E2)$ has to be modified as follows

In the approximate model the phase of the MMPP to Q (the composite queue) will be different from that of the MMPP to Q2 Hence in case 5, for computing $P[E2]$, We consider the chain of conditional events

$$E20 \{ \text{A cell arrives at an empty Q at time } t', J(t') = j' | X(\tau) = 0, J(\tau) = j \}$$

E21 { MMPP to Q2 is in phase j at τ }

E22 { c_2 cells arrive at Q2 in (t', t) and $J(t)=\ell$ | $X''(t')=0$, $J''(t')=j$ }

Using these events it can be verified that

$$P[E2] = \int_{t'=\tau}^t dU_{1j'}(t'-\tau) \theta''(j) P_{j\ell}''(c_2, t-t') \quad (6.4.4)$$

where $\theta''(j)$ is the steady state probability of finding the MMPP to Q2 in phase j at an arbitrary time instant. Let $X(t)$ and $J(t)$ denote the number of cells in the system Q and the phase of the MMPP to Q at time t , respectively.

The probability of occurrence of event 1 has to be evaluated for all the five cases under which $v''(t)$ is non-zero and the details are given in Appendix (6 C). Using (6 C 1)-(6 C 8) and the expressions for $P(E2)$ and $P(E3)$ corresponding to the exact model given in section (6.3) and Appendix (6 A) we get -

$$\begin{aligned} & P[0 < v''(t) < \sigma, J''(t)=\ell \mid X''(0)=i, J''(0)=j'', X'(0)=i', J'(0)=j'] \\ &= \sum_{c_1=1}^{\infty} \sum_{j=1}^N \sum_{c_2=0}^{\infty} \int_0^t d\Phi_{c_1 j}''(\tau) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1 D-c_2 D) \\ &+ (1-p_0') \sum_{j=1}^N \sum_{c_2=0}^{\infty} \int_0^t d\Phi_{0j}''(\tau) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-D-c_2 D) \\ &+ (1-p_0') \sum_{c_1=1}^{\infty} \sum_{j=1}^N \sum_{j'=1}^N \sum_{c_2=0}^{\infty} \int_0^{t-D} \left\{ \int_{\tau=\tau'+D}^t d\Phi_{0j'}''(\tau') dU_{c_1 j}''(\tau-\tau') \right\} \\ & \quad P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1 D-c_2 D) \Big] \\ &+ \sum_{c_1=1}^{\infty} \sum_{k=1}^M \sum_{j=1}^N \sum_{c_2=0}^{\infty} \int_0^t d\Phi_{c_1 k}'(\tau) y''(0, j) P_{j\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1 D-c_2 D) \end{aligned}$$

$$\begin{aligned}
& + \sum_{j=1}^{M'} \sum_{j=1}^N \sum_{c_2=0}^{\infty} \int_0^t d\Phi_{0j}^{(j)}(\tau) \sum_{k=1}^{M'} \int_0^{t-\tau} dU_{jk}^{(k)}(t-\tau-t') \theta''(j) P_j''(c_2, t') \\
& u(D-t') u(\sigma+t'-D-c_2D) \quad (6.4.5)
\end{aligned}$$

Comparing (6.4.5) with (6.2.10) it can be seen that the form of the equations are similar. The upper limits of j become N in (6.4.5) as the total number of phases of the MMPP to Q_2 is N . In (6.2.10) it is MN as the total number of phases of MMPP $\underline{1}$ and MMPP $\underline{2}$ as well as the composite MMPP to Q is MN . The upper limit for k in the 4th term of (6.4.5) is M as the total number of phases of MMPP $\underline{1}$ is M . The 5th term of (6.4.5) differs from that of (6.2.10) by the appearance of the factor $\theta''(j)$. The total number of phases of the MMPP to Q is denoted as M' .

In view of the similarity of (6.2.10) and (6.4.5), the application of the KRT and the computation of the LST of $W''(\sigma)$ does not require any additional effort. By applying the KRT to (6.4.5) and writing the resulting equation in matrix form we get

$$\begin{aligned}
W''(\sigma) &= \sum_{c_1=1}^{\infty} \xi^{*''} \sum_{c_2=0}^{\infty} x_{c_1}'' \int_0^D P''(c_2, t) u(\sigma+t-c_1D-c_2D) \\
&+ (1-p_0') \sum_{c_2=0}^{\infty} \xi^{*''} x_0'' \int_0^D P''(c_2, t) u(\sigma+t-D-c_2D) \\
&+ (1-p_0') \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*''} x_0'' \int_0^D \tilde{U}_{c_1}'' P''(c_2, t) u(\sigma+t-c_1D-c_2D) \\
&- (1-p_0') \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*''} x_0'' \int_0^{\infty} \int_0^D dU_{c_1}''(v) P''(c_2, t-v) u(D-t+v) u(\sigma+t-v-c_1D-c_2D)
\end{aligned}$$

$$\begin{aligned}
& + \sum_{c_1=1}^{\infty} \sum_{c_2=0}^{\infty} \xi^{*'} x'_{c_1} e y_0'' \int_{t=0}^D P''(c_2, t) u(\sigma + t - D - c_2 D) \\
& + \sum_{c_2=0}^{\infty} \int_{t=0}^D \xi^* x_0 \tilde{U}_1 e^{\theta} P''(c_2, t) u(\sigma + t - D - c_2 D) + W''(0)
\end{aligned} \quad (6.4.6)$$

Here, $W''(0)$ denotes the probability that $v''(t)$ is zero as $t \rightarrow \infty$ and $\tilde{U}_{c_1}'', \tilde{U}_1$ denote the LST of $U_{c_1}''(t)$ and $U_1(t)$ respectively evaluated at $s=0$. It may be noted that they are also equal to $U_{c_1}''(\infty)$ and $U_1(\infty)$ respectively and correspond to the approximate model. It can be noted the form of (6.4.6) differs from (6.2.18) only in the last term by the appearance of the term e^{θ} . In (6.4.6) the various parameters that are present correspond to the approximate model where as in (6.2.18) they correspond to the exact model. The LST of $W''(\sigma)$ denoted as $\tilde{W}''(s)$, is given by

$$\begin{aligned}
\tilde{W}''(s) &= W''(0) + \left[y_0'' [\Lambda'' - Q^*] [-\tilde{U}''(r, 0)] + W_c [\tilde{\Lambda}''(r, 0) - rI] + \mathcal{E}(s) \right] \left[R''(r) + sI \right]^{-1} \\
W_c &= (1 - p_0'') y_0'' (\Lambda'' - Q^*) + \frac{1}{x_0' e} y_0' (\Lambda' - Q^*) e (1 - x_0' e) y_0'' + y_0 \Lambda e^{\theta} - y_0'' \Lambda'' p_0' \quad (6.4.7)
\end{aligned}$$

$$\begin{aligned}
\mathcal{E}(s) &= (1 - p_0'') y_0'' (\Lambda'' - Q^*) \int_{v=0}^D \frac{1}{D} \int_{w=0}^v r P''(0, v-w) \Lambda'' dw e^{(w-v)R''(r)} e^{-vs} \\
&\quad \left\{ e^{[R''(r) + sI](D+v)} - I \right\} \quad (6.4.9)
\end{aligned}$$

6.5. AVERAGE QUEUEING DELAY OF Q2

The computation of the average queueing delay of a cell arriving at Q2 is straight forward. The average queueing delay at Q2 denoted as W_H , can be found by differentiating (6.3.41) w.r.t. s for the exact model. For the approximate model (6.4.7) should be used in place of (6.3.41). The computation of the moments of $\tilde{W}''(s)$ requires a lot of tedious manipulations. Let us consider a scenario in which the traffic offered at Q2 is expected to be small and typically the traffic offered (ρ) is less than about 0.3. In an ATM applica-

tion the service time/cell is of the order of few microsec with link capacity of 150 Bps and cell size of 53 bytes. Under these two conditions computation of the average queueing delays have manageable complexity and we shall assume these two conditions to be satisfied in this section. For the computation of the moments we consider two alternate approaches. In the first approach, the technique used to compute the moments of the queue lengths in Sec 5.4 is used to compute the moments of $\tilde{W}''(s)$. In the second approach we make use of the observation that the n th moment of queueing delay becomes negligible compared to its $(n-1)$ th moment when the traffic offered and service time/cell is small. We investigate the validity of this claim when we present more details of the second approach. The second approach requires less computational effort compared to the first approach. We consider more details of these approaches next.

First, the application of the results of Sec 5.4 on the computation of the moments of queue length for the computation of queueing delay at Q2, using the exact model, is considered. The extension of these results for the approximate model is indicated subsequently. To simplify the manipulations (6.3.45) is rewritten as follows

$$\tilde{W}''(s) [R''(r) + sI] = U(s) \quad (6.5.1)$$

where

$$U(s) = W''(0) [R''(r) + sI] + y_0'' [\underline{\Lambda}'' - \underline{Q}^*] [-\tilde{U}''(r, 0)] + W_c'' [\tilde{A}''(r, 0) - rI] + \mathcal{E}(s) \quad (6.5.2)$$

Substituting (3.6.1) in (6.5.1) and noting that $r = e^{-sD}$ we get

$$\tilde{W}''(s) \left[\underline{\Lambda}'' e^{-sD} - \underline{\Lambda}'' + \underline{Q}^* + sI \right] = U(s) \quad (6.5.3)$$

Next, We drop the superscripts of $''$ and \sim for simplicity of notation. The n th moments of $W(s)$ and $U(s)$ are denoted as $W^{(n)}(s)$ and $U^{(n)}(s)$ and their values at $s=0$ is denoted as $W^{(n)}$ and $U^{(n)}$ respectively. With this notation, differentiating (6.5.1) w.r.t. s we get

$$W^{(1)}(s) \left[\underline{\Lambda} e^{-sD} - \underline{\Lambda} + \underline{Q}^* + sI \right] + W(s) \left[-\underline{\Lambda} D e^{-sD} + I \right] = U^{(1)}(s) \quad (6.5.4)$$

$$W^{(2)}(s) \left[\underline{\Lambda} e^{-sD} - \underline{\Lambda} + \underline{Q}^* + sI \right] + 2W^{(1)}(s) \left[-\underline{\Lambda} D e^{-sD} + I \right] + W(s) \underline{\Lambda} D^2 e^{-sD} = U^{(2)}(s) \quad (6.5.5)$$

Multiplying (6.5.4) - (6.5.5) by \mathbf{e} , the $MN \times 1$ unity vector, evaluating at $s=0$ and noting that $\underline{Q}^* \mathbf{e}$ is a null vector, we get

$$W [I - \underline{\Lambda} D] \mathbf{e} = U^{(1)} \mathbf{e} \quad (6.5.6)$$

$$2W^{(1)} [I - \underline{\Lambda} D] \mathbf{e} + W \underline{\Lambda} D^2 \mathbf{e} = U^{(2)} \mathbf{e} \quad (6.5.7)$$

where $W(0)$ is denoted as W . Let π be the invariant vector of \underline{Q}^* . Adding $W^{(1)} \mathbf{e} \pi$ on both sides of (6.5.4) and evaluating at $s=0$ we get

$$W^{(1)} [\underline{Q}^* + \mathbf{e} \pi] + W [I - \underline{\Lambda} D] = U^{(1)} + W^{(1)} \mathbf{e} \pi \quad (6.5.8)$$

It may be verified (see for e.g. Neuts[2]) that $\underline{Q}^* + \mathbf{e} \pi$ is nonsingular and π , \mathbf{e} can be written as

$$\pi [\underline{Q}^* + \mathbf{e} \pi] = \pi = \pi [\underline{Q}^* + \mathbf{e} \pi]^{-1} \quad (6.5.9)$$

$$[\underline{Q}^* + \mathbf{e} \pi] \mathbf{e} = \mathbf{e} = [\underline{Q}^* + \mathbf{e} \pi]^{-1} \mathbf{e} \quad (6.5.10)$$

Using (6.5.9) in (6.5.8) we get-

$$W^{(1)} = W^{(1)} \mathbf{e} \pi [\underline{Q}^* + \mathbf{e} \pi]^{-1} + \left\{ U^{(1)} - W [I - \underline{\Lambda} D] \right\} [\underline{Q}^* + \mathbf{e} \pi]^{-1} \quad (6.5.11)$$

$$= W^{(1)} \mathbf{e} \pi + \left\{ U^{(1)} - W [I - \underline{\Lambda} D] \right\} [\underline{Q}^* + \mathbf{e} \pi]^{-1} \quad (6.5.12)$$

Substituting (6.5.12) in (6.5.7) we get

$$2 \left\{ W^{(1)} \mathbf{e} \pi + [U^{(1)} - W [I - \underline{\Lambda} D]] [\underline{Q}^* + \mathbf{e} \pi]^{-1} \right\} [I - \underline{\Lambda} D] \mathbf{e} = U^{(2)} \mathbf{e} - W \underline{\Lambda} D^2 \mathbf{e} \quad (6.5.13)$$

Expanding (6.5.13) and using (6.5.9) we get

$$2\left\{W^{(1)}e\pi e - W^{(1)}e\pi\Lambda eD\right\} = U^{(2)}e - W\Lambda D^2e \\ - 2\left[U^{(1)} - W\left(I - \Lambda D\right)\right]\left[\underline{Q}^* + e\pi\right]^{-1}\left[I - \Lambda D\right]e \quad (6.5.14)$$

It may be noted that $\pi\Lambda e$ gives the average arrival rate at Q2, averaged over all the phases of MMPP. D is service time/cell. Hence $\pi\Lambda eD$ denotes the traffic offered at Q2 and let it be denoted as ρ . i.e.

$$\rho = \pi\Lambda eD \quad (6.5.15)$$

Substituting (6.5.15) in (6.5.14) and rewriting we get

$$W^{(1)}e = \frac{1}{2(1-\rho)}\left\{U^{(2)}e - W\Lambda D^2e \right. \\ \left. - 2\left[U^{(1)} - W\left(I - \Lambda D\right)\right]\left[\underline{Q}^* + e\pi\right]^{-1}\left[I - \Lambda D\right]e\right\} \quad (6.5.16)$$

Using (6.5.10) in (6.5.6) we get

$$\left\{U^{(1)}e - W\left[I - \Lambda D\right]\right\}\left[\underline{Q}^* + e\pi\right]^{-1}e = 0 \quad (6.5.17)$$

Using (6.5.17) in (6.5.16) we get

$$W^{(1)}e = \frac{1}{2(1-\rho)}\left\{U^{(2)}e - W\Lambda D^2e + 2\left[U^{(1)} - W\left(I - \Lambda D\right)\right]\left[\underline{Q}^* + e\pi\right]^{-1}\Lambda De\right\} \quad (6.5.18)$$

The average queueing delay at Q2 denoted as W_H is finally given by-

$$W_H = -W^{(1)}e \quad (6.5.19)$$

For computing W_H , the moments of $U(s)$ remains to be evaluated and this is considered next.

Substituting (3.6.20) and (3.6.1) in (6.5.2) we get

$$U(s) = W''(0)\left[\Lambda''e^{-sD} - \Lambda'' + \underline{Q}^* + sI\right] - y_0''\Lambda''p_0'e^{-sD} - (1-p_0')\frac{1}{D}y_0''\Lambda''e^{-sD} \\ \int_0^D dw e^{[\Lambda''e^{-sD} - \Lambda'' + \underline{Q}^*]w} + W_c\left\{e^{[\Lambda''e^{-sD} - \Lambda'' + \underline{Q}^*]D} - e^{-sD}I\right\} + \mathcal{E}(s) \quad (6.5.20)$$

To simplify the notation we define a function $F(s, w)$ as follows

$$F(s, w) = e^{[\Lambda'' e^{-sD} - \Lambda'' + Q^*]w} \quad (6.5.21)$$

As noted in Sec 6.4 when the traffic offered at Q2 is small and the service/cell is of the order of microseconds, $\mathcal{E}(s)$ can be neglected. Neglecting $\mathcal{E}(s)$ and using (6.5.21) in (6.5.20) we get

$$\begin{aligned} U(s) &= W''(0)[Q^* - \Lambda'' + sI] + [W''(0)\Lambda'' - y_0''\Lambda''p_0']e^{-sD} \\ &\quad - y_0''\Lambda''(1-p_0') \frac{1}{D} e^{-sD} \int_0^D F(s, w)dw + W_c F(s, D) \end{aligned} \quad (6.5.22)$$

Differentiating (6.5.22) w.r.t. s successively we get

$$\begin{aligned} U^{(1)}(s) &= W'''(0) - D[W''(0)\Lambda'' - y_0''\Lambda''p_0']e^{-sD} - y_0''\Lambda''(1-p_0') \frac{1}{D} \\ &\quad \left[-De^{-sD} \int_0^D F(s, w)dw + e^{-sD} \int_0^D F^{(1)}(s, w)dw \right] + W_c F^{(1)}(s, D) \end{aligned} \quad (6.5.23)$$

$$\begin{aligned} U^{(2)}(s) &= -D^2[W'''(0)\Lambda'' - y_0''\Lambda''p_0']e^{-sD} - y_0''\Lambda'' \frac{(1-p_0')}{D} \left[D^2 e^{-sD} \int_0^D F(s, w)dw \right. \\ &\quad \left. - 2De^{-sD} \int_0^D F^{(1)}(s, w)dw + e^{-sD} \int_0^D F^{(2)}(s, w)dw \right] + W_c F^{(2)}(s, D) \end{aligned} \quad (6.5.24)$$

Next, expanding the RHS of (6.5.21) using Taylor series, differentiation of $F(s, w)$ w.r.t. s is carried out as follows

$$F(s, w) = e^{[\Lambda'' e^{-sD} - \Lambda'' + Q^*]w} = \sum_{n=0}^{\infty} \frac{w^n}{n!} [\Lambda'' e^{-sD} - \Lambda'' + Q^*]^n \quad (6.5.25)$$

$$F^{(1)}(s, w) = \sum_{n=1}^{\infty} \sum_{m=0}^{n-1} \frac{w^n}{n!} []^m (-D)\Lambda'' e^{-sD} [\Lambda'' e^{-sD} - \Lambda'' + Q^*]^{n-m-1} \quad (6.5.26)$$

$$\begin{aligned} F^{(2)}(s, w) &= 2 \sum_{n=2}^{\infty} \sum_{m=1}^{n-1} \sum_{k=0}^{m-1} \frac{w^n}{n!} D^2 e^{-2sD} []^k \Lambda'' []^{m-k-1} \Lambda'' [\Lambda'' e^{-sD} - \Lambda'' + Q^*]^{n-m-1} \\ &\quad + \sum_{n=1}^{\infty} \sum_{m=0}^{n-1} \frac{w^n}{n!} []^m D^2 \Lambda'' e^{-sD} [\Lambda'' e^{-sD} - \Lambda'' + Q^*]^{n-m-1} \end{aligned} \quad (6.5.27)$$

where [] has been used to denote $[\underline{\Lambda}''e^{-sD} - \underline{\Lambda}'' + \underline{Q}^*]$ for ease of writing Closed form expressions, free from infinite summations, for $F^{(1)}(0,w)$ and $F^{(2)}(0,w)$ are given in Fischer [3]. We shall not reproduce them here as they are lengthy and integration of these expressions w.r.t. w become quite tedious. For evaluating $U^{(n)}(s)$ at $s=0$ for $n = 1, 2$, (6.5.25) - (6.5.27) have to be integrated w.r.t. w at $s = 0$. The upper limit for the integration is $w=D$. Since D is of the order of 10^{-6} , in the summations of (6.5.26) and (6.5.27) the terms containing the higher order powers of D can be neglected. It may be noted here that for typical traffic offered at Q_2 in the range of 0.05 to 0.3 and D of the order of microsec the entries of the matrix $\underline{\Lambda}''$ is of the order of 10^5 . Hence the term $\underline{\Lambda}''D$ is of the order of few hundredths to few tenths. The rate of phase transition of the MMPPs are small and is typically few tens/sec. This can be verified from the phase transition rates given in Fig 5.9 - Fig 5.41 in Sec 5.10. Hence, the term \underline{Q}^*D is of the order of 10^{-4} . Expanding the infinite summations and neglecting terms containing higher order powers of D we get

$$F^{(1)}(0,w) \cong -(\underline{\Lambda}''D)w \quad (6.5.28)$$

$$F^{(2)}(0,w) \cong 2(\underline{\Lambda}''D)^2 \frac{w^2}{2!} + D^2 \underline{\Lambda}''w \quad (6.5.29)$$

Substituting (6.5.28) - (6.5.29) in (6.5.23) and (6.5.24) and integrating w.r.t. w , we get

$$U^{(1)}(0) = W''(0)[I - \underline{\Lambda}''D] + p'_0 y''_0 \underline{\Lambda}''D + (1-p'_0) y''_0 [\underline{\Lambda}''D \mathcal{J}_0 + \frac{1}{2} \underline{\Lambda}''^2 D^2] + W_c D(I - \underline{\Lambda}''D) \quad (6.5.30)$$

$$U^{(2)}(0) = \left\{ W''(0)\underline{\Lambda}'' - p'_0 y''_0 \underline{\Lambda}'' - (1-p'_0) y''_0 \left[\underline{\Lambda}'' \mathcal{J}_0 + \frac{3}{2} (\underline{\Lambda}''D)\underline{\Lambda}'' + \frac{1}{3} (\underline{\Lambda}''D)^2 \underline{\Lambda}'' \right] - W_c \left[I - \underline{\Lambda}''D - (\underline{\Lambda}''D)^2 \right] \right\} D^2 \quad (6.5.31)$$

where

$$g_0 = \int_0^D dw e^{Q^*w} = \epsilon \pi D + [e^{Q^*D} - 1][\epsilon \pi + Q^*]^{-1} \quad (6.5.32)$$

It may be noted that (6.5.32) is obtained using (4.5.12). We are finally left with the evaluation of $W''(0)$. It may be noted that the virtual waiting time for a Q2 cell is zero if both Q1 and Q2 are empty. Hence the j^{th} element of $W''(0)$ is equal to the probability that both Q1 and Q2 are empty and the MMPP \underline{z} is in phase j at an arbitrary time instant. The composite queue Q is empty whenever both Q1 and Q2 are empty. The phase of the composite MMPP is equal to that of MMPP \underline{z} at all time instant. Hence $W''(0)$ is given by

$$W''(0) = y_0 \quad (6.5.33)$$

This completes the set of equations required for the evaluation of the queueing delay at Q2.

Next we consider the computation of the queueing delay at Q2 using the second method. For the moment we shall assume $W''(0)$, $W^{(1)}$ and $W^{(2)}$ to be scalars. Alternately we can concentrate on one of the components of these vectors. When the traffic offered at Q2 is low and the service time/cell is of the order of microseconds, the probability that a cell arriving at Q2 receives service immediately is not insignificant and may actually be of the order of few multiples of 0.01 to 0.1. $W^{(1)}$ is expected to be of the order 10^{-6} . This can be verified as follows:

$$-W''^{(1)} = \sum_{n=1}^{\infty} nD P[QL=n] + w \quad (6.5.34)$$

where $P[QL=n]$ denotes the probability that a Q2 cell finds n cells in Q2 when it arrives at Q2. The value of w depends on whether a Q1 cell or a Q2 cell is undergoing service when the cell of interest arrives at Q2 and has a maximum value of $0.5D$. From the examples considered in Chapter 5, it can be concluded

that the probability that the queue length exceeds about 20 becomes negligibly small. Hence $-W^{(1)}$ is of the order of D (2.8 microsec). Similarly $W^{(2)}$ is given by

$$W^{(2)} = \sum_{n=1}^{\infty} (nD)^2 P[QL=n] + \gamma \quad (6.5.35)$$

where γ is a constant of the order of D^2 . With queue lengths of Q_2 typically less than 20, $W^{(2)}$ is typically of the order of few tens of D^2 . Computing $W^{(n)}$ for $n = 3, 4$ it can be concluded that the n th moment of $\tilde{W}(s)$ is significantly smaller compared to its $(n-1)$ th moment. This observation simplifies the computation of the moments of $\tilde{W}(s)$. Even though this observation is arrived at by considering the case where $\tilde{W}(s)$ is a scalar, it can be verified that it is also valid for the case where it is a vector. The computation of W_H can be carried out as follows. Differentiating (6.5.3) w.r.t. s successively we get-

$$\tilde{W}^{(1)}(s)[] + \tilde{W}(s)[-\underline{\Lambda}''D e^{-sD} + I] = U^{(1)}(s) \quad (6.5.36)$$

$$\tilde{W}^{(2)}(s)[] + 2\tilde{W}^{(1)}(s)[-\underline{\Lambda}''D e^{-sD} + I] + W(s)\underline{\Lambda}''D^2 e^{-sD} = U^{(2)}(s) \quad (6.5.37)$$

where $[]$ denotes $[\underline{\Lambda}''e^{-sD} - \underline{\Lambda}'' + Q^* + sI]$. Evaluating (6.5.37) at $s=0$ we get

$$W^{(2)}(0)Q^* + 2W^{(1)}(0)[-\underline{\Lambda}''D + I] + W\underline{\Lambda}''D^2 = U^{(2)} \quad (6.5.38)$$

As noted earlier, the entries of the matrix Q^* is of the order of tens at the maximum. Hence the 1st term of (6.5.38) is very small compared to the second term and hence W_H is given by

$$W_H = -W^{(1)}(0)e = \frac{1}{2} [W\underline{\Lambda}''D^2 - U^{(2)}][I - \underline{\Lambda}''D]^{-1}e \quad (6.5.39)$$

It may be noted that W is equal to π . This follows by noting that as $\sigma \rightarrow \infty$, $W(\sigma) \rightarrow \pi$. Substituting (6.5.31) in (6.5.39) and replacing W by π we get-

$$W_H = \frac{1}{2} \left\{ [\pi - W^{(1)}(0)] \underline{\Lambda}'' + p_0' y_0'' \underline{\Lambda}'' + (1-p_0') y_0'' \left[\underline{\Lambda}'' \mathcal{J}_0 + \frac{3}{2} (\underline{\Lambda}''D) \underline{\Lambda}'' + \frac{1}{3} (\underline{\Lambda}''D)^2 \underline{\Lambda}'' \right] + W_c \left[I - \underline{\Lambda}''D - (\underline{\Lambda}''D)^2 \right] \right\} D^2 [I - \underline{\Lambda}''D]^{-1}e \quad (6.5.40)$$

where \mathcal{J}_0 is given by (6.5.32). This completes the set of equations for the evaluation of W_H . These equations are also valid for the approximate model discussed in 6.4. The corresponding parameters given in Sec. 6.4 have to be used for this case. It may be noted that $W''(0)$ can also be evaluated using (6.5.36) by evaluating it at $s=0$. Noting that W is very large compared to $W^{(1)}(0)$ we get

$$W^{(1)}(0)[Q^*] + W(0)[-A''D + I] = U^{(1)}(0) \quad (6.5.41)$$

$$W(0)[I - A''D] = W[I - A''D] \cong U^{(1)}$$

$$\begin{aligned} &= W''(0)[I - A''D] + p'_0 y''_0 A''D + (1-p'_0) y''_0 [A''D \mathcal{J}_0 + \frac{1}{2} A''^2 D^2] \\ &\quad + W_c D(I - A''D) \end{aligned} \quad (6.5.42)$$

Noting that W is equal to π , the invariant vector of Q^* and using (6.5.42) we get

$$\begin{aligned} W''(0) = \pi - \left\{ p'_0 y''_0 A''D + (1-p'_0) y''_0 [A''D \mathcal{J}_0 + \frac{1}{2} A''^2 D^2] \right. \\ \left. + W_c D(I - A''D) \right\} [I - A''D]^{-1} \end{aligned} \quad (6.5.43)$$

6.6. AVERAGE QUEUEING DELAY AT Q1

The average virtual waiting time of a cell arriving at Q1, denoted as W_L , can be found by applying the M/G/1 conservation law [4] to MMPP/D/1 queues with non-preemptive priority. Let the traffic intensity originating from the high, low priority classes be ρ'' , ρ' and the average queueing delay for the cells from these classes be denoted as W_H and W_L respectively. As the average unfinished work as well as the average residual service time, averaged over all the priority classes, are constant for a non-preemptive priority system the sum $\rho''W_H + \rho'W_L$ is independent of the order in which the cells from

various classes are served. In particular let the total traffic originating from both the classes be ρ . As in Sec. 6.1, let us consider a composite queue Q to which the traffic originating from both the classes are fed and are served on a FCFS basis. Let the average queueing delay of a cell arriving at Q be W_T , then by the application of the conservation law for the non-preemptive priority system we get-

$$\rho W_T = \rho'' W_H + \rho' W_L \quad (6.6.1)$$

The average queueing delay of a cell at the composite queue Q can be shown (see for e.g. Fischer [3], Heffes [5]) to be given by-

$$W_T = \frac{1}{2(1-\rho)} \left\{ \pi \underline{\Lambda} D^2 e + 2 \left[y_0 - \pi \left(I - \underline{\Lambda} D \right) \right] \left[Q^* + e\pi \right]^{-1} \left[I - \underline{\Lambda} D \right] e \right\} \quad (6.6.2)$$

$$\underline{\Lambda} = \Lambda' \otimes I_N + I_M \otimes \Lambda'' \quad (6.6.3)$$

$$\rho = \pi \underline{\Lambda} e D \quad (6.6.4)$$

where $\underline{\Lambda}$ is the arrival rate matrix corresponding to the composite arrival process modelled again as an MMPP and π is the equilibrium probability vector of Q^* , the infinitesimal generator matrix of the composite process. The traffic intensity ρ'' corresponding to Q_2 is given by-

$$\rho'' = \pi \underline{\Lambda}'' e D \quad (6.6.5)$$

ρ' is obtained by replacing $\underline{\Lambda}''$ by $\underline{\Lambda}'$ in (6.6.5). It may be noted that the infinitesimal generator matrix is the same for the MMPPs MMPP 1, MMPP 2 and the composite MMPP to Q . Hence the invariant vector is π for all the three MMPPs. For the approximate model the infinitesimal generator matrix of the MMPPs corresponding to all the three queues are different and hence the invariant vectors of the composite MMPP and MMPP 2 should be used in (6.6.4) and (6.6.5) respectively. A corresponding change is also required for the computation of ρ' . Computing W_T and W_H using (6.6.2) and (6.4.40) and substituting in (6.6.1), W_L , the average queueing delay of a low priority cell arriving at Q_1 can be computed.

For analytical simplicity, we have assumed the low and high priority cells to arrive at two separate queues. However, it is obvious that as long as the service discipline is the same (a single server is shared between the two classes on a non preemptive basis), the average queueing delay experienced by each priority class is the same irrespective of whether they arrive at two separate queues or at a single queue where the high priority cells are given Head of the Line (HOL) non preemptive priority for service.

6.7. EXTENSION FOR M/D/1 QUEUES WITH NON PREEMPTIVE PRIORITY

The results of Sec 6.2-6.4 are applied here for the special case where the arrivals at both Q1 and Q2 are individually modelled as two Poisson processes. In particular, an expression for the LST of the cdf of the virtual waiting time at the higher priority queue Q2 and expressions for the average virtual virtual waiting time at Q1 as well as at Q2 are obtained. For an M/D/1 queue with non preemptive priority, the expressions for these characteristics have been obtained using an alternate method and are given in Conway [6], Gravey [7]. Comparison of the results obtained using these alternate approaches is used as an additional validation of the method suggested for the MMPP/D/1 system. Using the method proposed in [7], $W_2(s)$, the LST of the cdf of the virtual waiting time at Q2, is given by

$$W_2(s) = \frac{(1-\rho)s + \lambda'(1-e^{-sD})}{s - \lambda''(1-e^{-sD})} \quad (6.7.1)$$

$$W_2(0) = (1 - \rho) \quad (6.7.2)$$

$$\rho = (\lambda' + \lambda'')D \quad (6.7.3)$$

where λ' , λ'' are the cell arrival rates to Q1 and Q2 respectively. Using (6.7.1), the average virtual waiting time at Q2, denoted as $\overline{W_2}$, can be obtained as

$$\overline{W_2} = \frac{\lambda''}{2[1-\lambda''D]} \left[1 + \frac{\lambda'}{\lambda''} \right] \quad (6.7.4)$$

Next the application of the results of Sec. 6.2-6.3 are considered. For the special case of Poisson arrivals Λ 's and Q^* 's are scalars and Q^* , Q^{**} , Q^{***} are zero. Denoting the scalar arrival rates to Q , Q_1 , Q_2 as λ , λ' and λ'' respectively and using (3.5.20) we get

$$\tilde{U}''(r,0) = p_0' e^{-sD} + (1-p_0') e^{-sD} \left[D(\tilde{e}^{-sD} - 1) \right]^{-1} \left[\tilde{A}''(r,0) - 1 \right] \quad (6.7.5)$$

$$\tilde{A}''(r,0) = e^{-\lambda''(e^{-sD} - 1)D} \quad (6.7.6)$$

$$R''(r) = \lambda''(e^{-sD} - 1) \quad (6.7.7)$$

$$\pi = \pi' = \pi'' = 1 \quad (6.7.8)$$

$$Q^* = Q^{**} = Q^{***} = 0 \quad (6.7.9)$$

Substituting equations (6.7.5)-(6.7.9) in (6.3.41) the LST of the cdf of the virtual waiting time and W_H , the average virtual waiting time at Q_2 can be obtained. W_H is given by

$$W_H = - \frac{d}{ds} [\tilde{W}''(s)] \Big|_{s=0} \quad (6.7.10)$$

$$= \left\{ (1-W''(0)) + p_0' y_0'' + (1-p_0') y_0'' \left[1 + \frac{3}{2} (\lambda''D) + \frac{1}{3} (\lambda''D)^2 \right] + W_c \left[1 - \lambda''D - (\lambda''D)^2 \right] \right\} \frac{\lambda''D^2}{2} [1 - \lambda''D]^{-1} \quad (6.7.11)$$

$$W''(0) = 1 - \left[y_0'' \lambda'' D \left\{ 1 + (1-p_0') \frac{\lambda''D}{2} \right\} + W_c D (1 - \lambda''D) \right] [1 - \lambda''D]^{-1} \quad (6.7.12)$$

$$W_c = (1-2p_0') y_0'' \lambda'' + (1-y_0') y_0'' \lambda' + y_0 \lambda \quad (6.7.13)$$

In this case \mathcal{J}_0 is a scalar and is equal to D . Substituting (6.7.11) in

(6.6.1) W_L can be found

6.8. NUMERICAL RESULTS

The average virtual waiting time at Q1 and Q2 are computed using both the exact model and the approximate model for some typical mixes of low and high priority traffic and the results are presented here. As in Chapter 5, we denote the exact and the approximate models as model I and II respectively. The computed results corresponding to MMPP arrivals are also compared with those obtained using simulation. For the non-preemptive M/D/1 priority system with two priority classes, the average queueing delays at Q2 computed using the two alternate approaches mentioned in Sec 6.7 are presented.

The traffic to Q1 and Q2 are assumed to originate from N1, N2 constant bit rate on/off sources with on bit rates of 1 Mbps, average on duration of 33 msec and percentage on duration of 35%. An output link capacity of 150 Mbps and cell size of 53 bytes are assumed. The composite traffic to Q1 from N1 sources is approximated by a 2 phase MMPP using the method proposed in [7]. Similarly the traffic from N2 sources fed to Q2 is also approximated by 2 phase MMPP. The parameters (Λ', Q^{**}) , (Λ'', Q^{**}) are thus found. Knowing these parameters, the parameters of the exact model (viz $\underline{\Lambda}'$, $\underline{\Lambda}''$, \underline{Q}^*) as well as that of the MMPP to Q are found.

Using the parameters of the MMPPs to Q1 and Q2, the simulation routine discussed in Sec 5.9 is used to obtain the average virtual waiting times at queues Q1 and Q2 (equations (5.9.8) and (5.9.9) are used for this purpose). For some typical values of N1 and N2, the average delays at Q2 and Q1 obtained using simulation are presented in Fig 6.1 and Fig 6.2 respectively. The normalized average delay (queueing delay + service time normalized by D, the service time) is shown as a function of the low priority traffic offered. In these figures, ρ'' denotes the average load offered to the high priority queue.

Next, for the exact model, the numerical procedure given in Sec 5.9 is

used to find y_0'' , y_0' and y_0 . Substituting these in (6.6.1)-(6.6.5), the average virtual waiting time at Q1 and Q2 are found and the results are shown in Fig 6.1 - Fig 6.3. From Fig 6.1 and Fig 6.2 it can be concluded that the results obtained using the exact model (model I) agree well with those of simulation for both the low and high priority queues.

For the approximate model (model II) the average queueing delays at Q2 are computed using (6.6.1)-(6.6.5) by replacing the parameters of the exact model by those of model II. The normalized delays are again computed and compared with those of the exact model in Fig 6.3. From this figure it can be concluded that the average delays obtained using both the exact model and the approximate model match well. We examine why the approximate model which is not completely satisfactory for the computation of the QLDs at Q1 and Q2 gives accurate results for queueing delays. From (6.4.40) it can be noted that for the computation of W_H in addition to the parameters of the arrival process only the parameters y_0'' , y_0' and y_0 need to be known. As noted in Sec 5.10 p_0' and p_0'' computed using both the exact and the approximate models are only marginally different. Hence the average delays computed using models I and II match well.

As an additional validation of the results obtained here, the queueing delays for the degenerate case of M/D/1 system with two priority classes and non-preemptive priority discipline are evaluated. The average queueing delays at Q2 as predicted by the method given in [7] is computed using (6.7.4) for some typical combinations of low and high priority traffic. Alternately, by treating the M/D/1 queue as a degenerate case of MMPP/D/1 queue the delays at Q2 is computed using (6.7.10)-(6.7.13). The results obtained using both the methods are shown in Fig 6.4. As in Fig 6.1- Fig 6.3, the normalized delays are shown in Fig 6.4 and ρ'' denotes the traffic offered at the high

priority queue. It can be seen that the results obtained using both the approaches match well. Since the average delays at the high priority queue match, using (6.6.1), it can be concluded that the low priority delays also match.

APPENDIX (6 A)

In this appendix, for the exact model, we compute the probability of occurrence of the events E1, E2 and E3 corresponding to the cases 2-5 referred to in section (6.2). We compute the probability of these events for each case separately as follows.

Case 2 A Q2 cell departed at τ , leaving Q2 empty and Q1 non empty. In this case $c_1 = 0$ and $c_2 \geq 0$ and $c_1' > 0$. $P[E1]$, in this case can be obtained by multiplying (6.2.4) evaluated at $c_1=0$ by the probability that $X'(\tau) > 0$ (denoted as $1-p_0'$). $P[E2]$, $P[E3]$ corresponding to this case can be obtained using (6.2.5) and (6.2.8) with $c_1=0$. Multiplying these expressions for $P(E1)$, $P(E2)$ and $P(E3)$, integrating over τ and removing the conditions on c_2 and $J(\tau)$ we get the 2nd term of (6.2.10).

Case 3 A Q1 cell departed at τ , $X''(\tau) > 0$ and $X'(\tau) \geq 0$. In this case $c_1 > 0$ and $c_2 \geq 0$ and $c_1' \geq 0$. This case implies that at $\tau-D$, Q2 was empty and Q1 non empty. $P[Q1 \text{ non-empty at } \tau-D]$ is given by $1-p_0'$. Let the time of the latest departure from Q2 before time t be τ' . Obviously in the interval (τ', τ) Q2 does not receive any service and the busy period of Q2 starts again at τ . $P(E1)$ can be obtained as the product of the probabilities of the following three events E10 {Q1 not empty at $\tau-D$ },

$$E11 \{X''(\tau')=0, J(\tau')=j \mid X''(0)=1 \text{ and } J(0)=j'\}$$

$$E12 \{X''(\tau)=c_1, J(\tau)=j \mid X''(\tau')=0, J(\tau')=j'\}$$

The computation of the probability of occurrence of these events is straight

forward and are given by-

$$P(E10) = (1-p'_0) \quad (6 A 1)$$

$$P(E11) = d\phi_{0J}^{1J'}(\tau') \quad (6 A 2)$$

$$P(E12) = dU_{c1}''(\tau-\tau') \quad (6 A 3)$$

$U_k''(t)$ has been defined immediately after (6 2 10) and is related to $U_k''(t)$ defined in section 3 3 as follows

$$U_k''(t) = U_k''(t) \Big|_{Q1 \text{ not empty}} \quad (6 A 4)$$

The RHS of (6 A 4) can be computed using (3 4 6) $P(E2)$ and $P(E3)$ for this case are the same as that for case 1 Using (6 A 1)-(6 A 3) we get the terms inside the square bracket of the third term of (6 2 10) In this term the upper limit for τ' is $t-D$ and can be verified as follows The interval (τ', t) consists of two parts service time for $Q1$ cells and service received at time t by a $Q2$ cell which began service at τ τ' is maximum if these two parts are the minimum The maximum τ' occurs when the time taken for these two parts are equal to D and 0 respectively, i.e. between τ' and t one $Q1$ cell receives service and just when $Q2$ starts receiving service at time t , the cell of interest arrives The lower limit for τ is equal to $\tau'+D$ and occurs when the BP of $Q2$ starts again after 1 $Q1$ cell service

Case 4 A $Q1$ cell departed at τ , $X'(\tau) > 0$ and $X''(\tau) = 0$ In this case $c_1 = 0$ and $c_2 \geq 0$ and $c_1' > 0$ Considering the two independent events $E11$ $\{X'(\tau) = 1' (1' > 0) \mid X'(0) = 1'', J'(0) = j''\}$ and $E12$ $\{X''(\tau) = 0, \tau \text{ an arbitrary time instant, } j(\tau) = j \mid X''(0) = 1, j(0) = j'\}$, $P(E1)$ can be obtained It can be shown that

$$P(E11) = \sum_{j=1}^{MN} P[X'(\tau)=c1, J'(\tau)=j | X'(0)=i'', J'(0)=j''] = \sum_{j=1}^{M \setminus} d\phi'_{c1j}{}^{i''j''}(\tau) \quad (6 A 5)$$

$$P(E12) = y''(0, j) \quad (6 A 6)$$

where $y''(0, j)$ denotes the probability that Q2 is empty at an arbitrary time instant with $\underline{j}(0) = j$ $\phi'_{c1j}{}^{i''j''}(\tau)$ is defined for Q1 similar to that for Q2 $P(E1)$ and $P(E3)$ are the same as that for case 1. Multiplying (6 2 5), (6 A 5)-(6 A 6) and (6 2 8) and removing the conditions on $c1$, τ , j and we get the 4th term of (6 2 10)

Case 5 Either a Q1 or a Q2 cell departed at τ and Q2 as well as Q1 are empty at τ . In this case $c1 = 0$ and $c2 \geq 0$ and $c1' = 0$. For the evaluation of $P[E1]$ and $P[E2]$ case we consider a third queue Q to which the low as well as high priority cells are fed and served on a FCFS basis. It can be seen that whenever Q1 and Q2 are empty simultaneously Q is also empty. Hence $P[E1]$ can be found using this queue as

$$P(E1) = P[X(\tau)=0, \underline{j}(\tau)=j | X(0)=i'', \underline{j}(0)=j''] = d\phi_{0j}{}^{i''j''}(\tau) \quad (6 A 7)$$

For computing $P[E2]$, We consider the chain of conditional events E21 {A cell arrives at an empty Q at time t' , $\underline{j}(t')=j | X(\tau)=0, \underline{j}(\tau)=j$ }, E22 { $c2$ cells arrive at Q2 in (t', t) and $\underline{j}(t)=\ell | X''(t')=0, \underline{j}(t')=j$ }. It can be shown that

$$P[E2] = \int_{t'=\tau}^t dU_{1j} (t'-\tau) P''_{j\ell}(c2, t-t') \quad (6 A 8)$$

Using a change of variable of $t = t' - \tau$, (6 A 8) can be rewritten as

$$P[E2] = \int_{t=0}^{t-\tau} dU_{1j} (t-t-\tau) P''_{j\ell}(c2, t) \quad (6 A 9)$$

Changing the dummy variable of t to t' in (6 A 9) we get

$$P[E2] = \int_{t'=0}^{t-\tau} dU_{1J} (t-t'-\tau) P_{J\ell}''(c_2, t') \quad (6 A 10)$$

$P(E3)$ is similar to that for case 1 with the following modification. The cell undergoing service at t began service only at t' . Using this, we get

$$P[E3] = u(\sigma + t - \tau - t' - D - c_2 D) \quad (6 A 11)$$

Using (6 A 7), (6 A 10) and (6 A 11) and proceeding as for the previous cases we get the last term of (6 2 10)

APPENDIX (6.B)

In this appendix, the application of the key renewal theorem to the terms (2-5) of the RHS of equation (6 2 10) and the simplification of the resulting terms are considered.

Application of the KRT to the 2nd and the 4th terms of (6 2 10) is straight forward and can be carried out along the same lines as for the first term. To apply the KRT to the third term of (6 2 10) we first choose a change of variable of $v = \tau - \tau'$. When $\tau = \tau' + D$, $v = D$ and $\tau = t$, $\Rightarrow v = t - \tau'$. Let the third term of (6 2 10) obtained by leaving out the summations and constant terms be denoted as $T3$. With the change of variable, $T3$ becomes

$$\begin{aligned} T3 &= \int_{\tau'=0}^{t-D} \int_{\tau=\tau'+D}^t d\Phi_{0J}''(\tau') dU_{c1}''(\tau-\tau') P_{J\ell}''(c_2, t-\tau) u(D-t+\tau) u(\sigma+t-\tau-c_1 D-c_2 D) \\ &= \int_{\tau'=0}^{t-D} \int_{v=D}^{t-\tau'} d\Phi_{0J}''(\tau') dU_{c1}''(v) P_{J\ell}''(c_2, t-\tau'-v) u(D-t+\tau'+v) u(\sigma+t-\tau'-v-c_1 D-c_2 D) \end{aligned} \quad (6 B 1)$$

Applying the KRT now and denoting the resulting expression as $\mathcal{T}3$ we get

$$\lim_{t \rightarrow \infty} T3 = \mathcal{T}3 = \frac{1}{m''(0, J)} \int_{t=0}^{\infty} \int_{v=D}^t dU_{c1}''(v) P_{J\ell}''(c_2, t-v) u(D-t+v) u(\sigma+t-v-c_1 D-c_2 D) \quad (6.B 2)$$

To simplify $\mathcal{J}3$ further, a change of variable of $v' = t - v$ is used. When $v = D$, $v' = t - D$ and $v = t$, $\Rightarrow v' = 0$. With this modification, $\mathcal{J}3$ becomes

$$\mathcal{J}3 = \frac{1}{m''(0, t)} \int_{t=0}^{\infty} \int_{v'=0}^{t-D} dU_{c1}''(t-v') P_{j\ell}''(c2, v') u(D-v') u(\sigma+v'-c1D-c2D) \quad (6 B 3)$$

The upper limit of v' in (6 B 10) can be changed to be t by adding and subtracting the missing portion of the integral in the interval $(t-D, t)$ and (6 B 3) becomes

$$\begin{aligned} \mathcal{J}3 &= \frac{1}{m''(0, t)} \int_{t=0}^{\infty} \int_{v'=0}^t dU_{c1}''(t-v') P_{j\ell}''(c2, v') u(D-v') u(\sigma+v'-c1D-c2D) \\ &\quad - \frac{1}{m''(0, t)} \int_{t=0}^{\infty} \int_{v'=t-D}^t dU_{c1}''(t-v') P_{j\ell}''(c2, v') u(D-v') u(\sigma+v'-c1D-c2D) \end{aligned} \quad (6 B 4)$$

Let the first, second term on the RHS of (6 B 4) be denoted $\mathcal{J}31$ and $\mathcal{J}32$ respectively. Next the simplification of $\mathcal{J}31$ is considered. The upper limit on v' in $\mathcal{J}31$ can be changed to be t as

$$dU_{c1}''(t-v') = 0 \text{ for } v' > t \quad (6 B 5)$$

As the upper limits on both t and v' are now ∞ , we next interchange the order of the integrals w.r.t. t and v' . Further, we use a change of variable of $t' = t - v'$. When $t = 0$, $t' = -v'$ and $t = \infty \Rightarrow t' = \infty$. With this change $\mathcal{J}31$ can be written as-

$$\mathcal{J}31 = \frac{1}{m''(0, t)} \int_{v'=0}^{\infty} \int_{t=0}^{\infty} dU_{c1}''(t-v') P_{j\ell}''(c2, v') u(D-v') u(\sigma+v'-c1D-c2D) \quad (6 B 6)$$

$$= \frac{1}{m''(0, t)} \int_{v'=0}^{\infty} \int_{t'=-v'}^{\infty} dU_{c1}''(t') P_{j\ell}''(c2, v') u(D-v') u(\sigma+v'-c1D-c2D) \quad (6 B 7)$$

$$\begin{aligned}
&= \frac{1}{m''(0, t)} \int_{v'=0}^{\infty} \int_{t'=-v'}^0 dU_{c_1}''(t') P_{j\ell}''(c_2, v') u(D-v') u(\sigma+v'-c_1D-c_2D) \\
&+ \frac{1}{m''(0, t)} \int_{v'=0}^{\infty} \int_{t'=0}^{\infty} dU_{c_1}''(t') P_{j\ell}''(c_2, v') u(D-v') u(\sigma+v'-c_1D-c_2D) \quad (6 B 8)
\end{aligned}$$

$$= \frac{1}{m''(0, t)} \int_{v'=0}^D U_{c_1}''(\infty) P_{j\ell}''(c_2, v') u(\sigma+v'-c_1D-c_2D) \quad (6 B 9)$$

(6 B 9) is obtained from (6 B 8) by noting that the first term on the RHS of (6 B 8) is zero in view of (6 B 5). The upper limit on v' is changed to be D as the term $u(D-v')$ is zero for v' greater than D . Next the simplification of \mathcal{T}_{32} is considered. Using a change of variable of $v = t-v'$ we get

$$\begin{aligned}
\mathcal{T}_{32} &= \frac{1}{m''(0, t)} \int_{t=0}^{\infty} \int_{v'=t-D}^t dU_{c_1}''(t-v') P_{j\ell}''(c_2, v') u(D-v') u(\sigma+v'-c_1D-c_2D) \\
&= \frac{1}{m''(0, t)} \int_{t=0}^{\infty} \int_{v=0}^D dU_{c_1}''(v) P_{j\ell}''(c_2, t-v) u(D-t+v) u(\sigma+t-v-c_1D-c_2D) \quad (6 B 10)
\end{aligned}$$

Next, the application of the KRT to the fifth term of (6.2.10) is considered. Let the 5th term of (6.2.10) leaving out the constant and the summations be denoted as T_5

$$T_5 = \int_{\tau=0}^t d\Phi_{0j}''(\tau) \int_{t'=0}^{t-\tau} dU_1(t-\tau-t') P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-c_2D-D) \quad (6 B 11)$$

Let $\mathcal{T}_5 = \lim_{t \rightarrow \infty} T_5$. Applying the KRT to \mathcal{T}_5 we get

$$\mathcal{T}_5 = \frac{1}{m(0, t)} \int_{t=0}^{\infty} \int_{t'=0}^t dU_1(t-t') P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-c_2D-D) \quad (6 B 12)$$

In \mathcal{T}_5 , the upper limit on t' can be changed to be ∞ as

$$dU_1(t-t', c_2) = 0 \text{ for } t' > t \quad (6 B 13)$$

As the limits on t and t' in 75 are now ∞ , the order of integration can be changed. Further using a change of variable of $v=t-t'$ we get

$$75 = \frac{1}{m(0, j)} \int_{t'=0}^{\infty} \int_{t=0}^{\infty} dU_1(t-t') P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-c_2 D-D) \quad (6 B 14)$$

$$= \frac{1}{m(0, j)} \int_{t'=0}^{\infty} \int_{v=-t'}^{\infty} dU_1(v) P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-D-c_2 D) \quad (6 B 14)$$

$$= \frac{1}{m(0, j)} \int_{t'=0}^{\infty} \int_{v=-t'}^0 dU_1(v) P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-D-c_2 D) \\ + \frac{1}{m(0, j)} \int_{t'=0}^{\infty} \int_{v=0}^{\infty} dU_1(v) P_{j\ell}''(c_2, t') u(D-t') u(\sigma+t'-D-c_2 D) \quad (6 B 15)$$

$$= \frac{1}{m(0, j)} \int_{t'=0}^D U_1(\infty) P_{j\ell}''(c_2, t') u(\sigma+t'-D-c_2 D) \quad (6 B 16)$$

(6 B 16) is obtained from (6 B 15) by noting that the 1st term of (6 B 15) is zero in view of (6 B 13). The upper limit of t' is changed to be D as $u(D-t')$ is zero for t' greater than D .

APPENDIX (6.C)

In this appendix, the details on the computation of $P(E1)$ corresponding to each of the 5 cases under which the virtual waiting time at Q2 is non-zero, are presented for the approximate model. It may be recalled that event 1 denotes the following conditional event

$$E1 \{ X''(\tau)=c_1, X'(\tau)=c_1', J''(\tau)=j | X''(0)=1, J''(0)=j, X'(0)=1, J'(0)=j' \}$$

Case 1 A Q2 cell departs at τ and leaves Q2 non empty. In this case $c_1 > 0$ and $c_2 \geq 0$ and $c_1' \geq 0$. The condition $c_1' \geq 0$ implies that $X'(\tau) > 0$.

turn implies that $P[X'(\tau) \geq 0 | X'(0) = 1] = 1$ and $P(E1)$ is independent of $X'(\tau)$

Hence $P(E1)$ can be computed by considering the event

$\{X''(\tau) = c_1 (c_1 > 0), J''(\tau) = j \mid X''(0) = 1, J''(0) = j\}$ and is given by

$$P(E1) \Big|_{\text{Case 1}} = d\phi_{c_1 j}^{1j''}(\tau) \quad (6 C 1)$$

where $\phi_{c_1 j}^{1j''}(\tau)$ is defined in the same manner as the exact model. In this case $Q''()$ corresponds to the approximate model.

Case 2 A Q2 cell departed at τ , leaving Q2 empty and Q1 non empty. In this case $c_1 = 0$ and $c_2 \geq 0$ and $c_1' > 0$. $P(E1)$, in this case can be obtained by multiplying (6 C 1) evaluated at $c_1 = 0$ by the probability that $X'(\tau) > 0$ (denoted as $1 - p_0'$).

Case 3 A Q1 cell departed at τ , $X''(\tau) > 0$ and $X'(\tau) \geq 0$. In this case $c_1 > 0$ and $c_2 \geq 0$ and $c_1' \geq 0$. This case implies that at $\tau - D$, Q2 was empty and Q1 non empty. $P[Q1 \text{ non-empty at } \tau - D]$ is given by $1 - p_0'$. Let the time of the latest departure from Q2 before time t be τ' . Obviously in the interval (τ', τ) Q2 does not receive any service and the busy period of Q2 starts again at τ . $P(E1)$ can be obtained as the product of the probabilities of the following three events: E10 {Q1 not empty at $\tau - D$ },

E11 $\{X''(\tau') = 0, J''(\tau') = j \mid X''(0) = 1 \text{ and } J''(0) = j\}$

E12 $\{X''(\tau) = c_1, J''(\tau) = j \mid X''(\tau') = 0, J''(\tau') = j\}$

The computation of the probability of occurrence of these events is straightforward and are given by-

$$P(E10) = (1 - p_0') \quad (6 C 2)$$

$$P(E11) = d\phi_{0j}^{1j''}(\tau') \quad (6 C 3)$$

$$P(E12) = dU_{c_1 j}''(\tau - \tau') \quad (6 C 4)$$

$U_k''(t)$ is defined in (6 A 4) and in this case corresponds to the approximate

model

Case 4 A Q1 cell departed at τ , $X'(\tau) > 0$ and $X''(\tau) = 0$. In this case $c_1 = 0$ and $c_2 \geq 0$ and $c_1' > 0$. Considering the two independent events E11 $\{X'(\tau) = i' (i' > 0) \mid X'(0) = i'', J'(0) = j'\}$ and E12 $\{X''(\tau) = 0, \tau \text{ an arbitrary time instant}, J''(\tau) = j \mid X''(0) = i, J''(0) = j''\}$, $P(E1)$ can be obtained. It can be shown that-

$$P(E11) = \sum_{j=1}^{MN} P[X'(\tau) = c_1, J'(\tau) = j \mid X'(0) = i'', J'(0) = j'] = \sum_{j=1}^{MN} d\phi_{c_1 j}^{i'' j'}(\tau) \quad (6 C 5)$$

$$P(E12) = y''(0, j) \quad (6 C 6)$$

where $y''(0, j)$ denotes the probability that Q2 is empty at an arbitrary time instant with $J(0) = j$. $\phi_{c_1 j}^{i'' j'}(\tau)$ is defined for Q1 similar to that for Q2 corresponding to the approximate model. $P(E1)$ and $P(E3)$ are the same as that for case 1.

Case 5 Either a Q1 or a Q2 cell departed at τ and Q2 as well as Q1 are empty at τ . In this case $c_1 = 0$ and $c_2 \geq 0$ and $c_1' = 0$. For the evaluation of $P(E1)$ we consider a third queue Q to which the low as well as high priority cells are fed and served on a FCFS basis. It can be seen that whenever Q1 and Q2 are empty simultaneously Q is also empty. Hence $P(E1)$ can be found using this queue as

$$P(E1) = P[X(\tau) = 0, J(\tau) = j \mid X(0) = i'', J(0) = j''] = d\phi_{0j}^{i'' j''}(\tau) \quad (6 C 7)$$

where $J(t)$ is the phase of the MMPP to Q. $\phi_{0j}^{i'' j''}(\tau)$ is obtained by replacing the parameters of Q2 by those of Q corresponding to the approximate model.

REFERENCES

- 1 V Ramaswami, "The N/G/1 queue and its detailed analysis," *Adv Appl Prob*, Vol.12, pp 222-261, Mar 1980
- 2 M F Neuts, "Structured Stochastic matrices of the M/G/1 type and their applications", Marcel Dekker, New York, 1989
- 3 W Fischer, K Meier-Hellstern, "The Markov Modulated Poisson Process (MMPP) cookbook", *Performance Evaluation*, 18, 1993, pp 149-171
- 4 L Kleinrock, "Communication Nets Stochastic Message Flow and Delay", McGraw-Hill, New York, 1964
- 5 H Heffes and D M Lucantoni, "A Markov Modulated characterization of Packetized voice and Data traffic and Related Statistical Multiplexer Performance", *IEEE J SAC*, No 6, pp 856- 867, Sep 1986
- 6 R W Conway, W L Maxwell and L W Miller, "Theory of Scheduling", Addison - Wesley publishing company, Massachusetts, 1967
- 7 A Gravey and G Hebuterne, "Mixing time and loss priorities in a single server queue", *Proc ITC -13*, Copenhagen, pp 147-152, June 1991

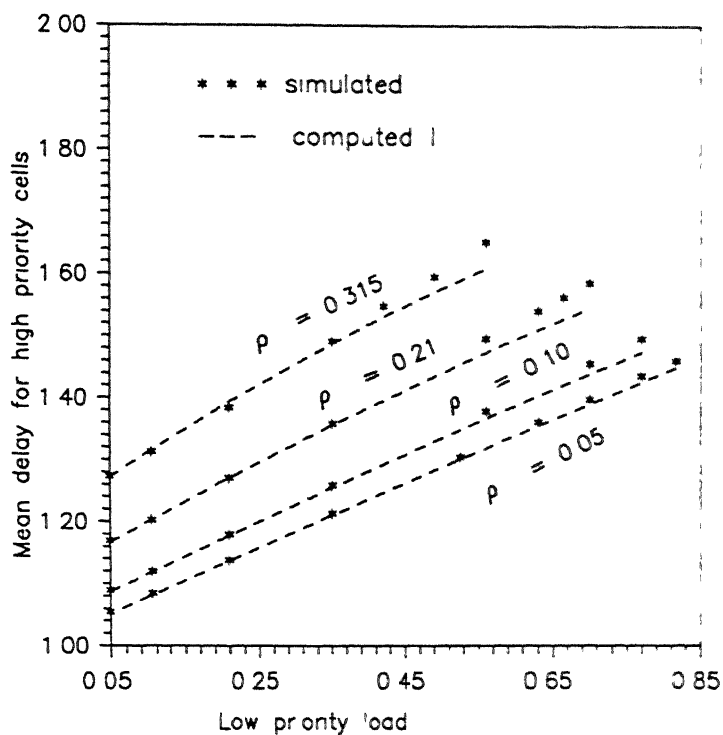


Fig 6.1 Mean delay for Q2 cells in an MMPP/D/1 priority system obtained using simulation and computation model

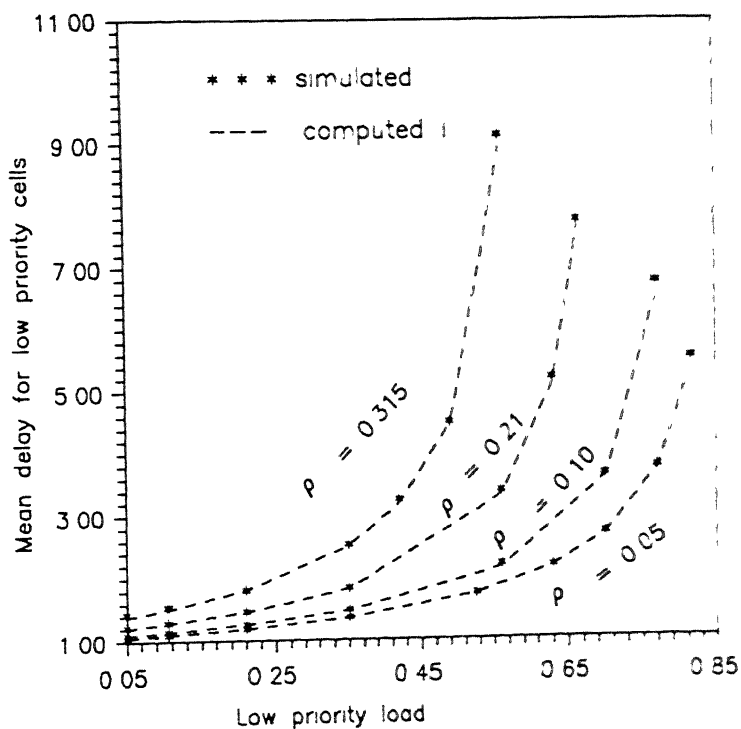


Fig 6.2 Mean delay for Q1 cells in an MMPP/D/1 priority system obtained using model I and simulation

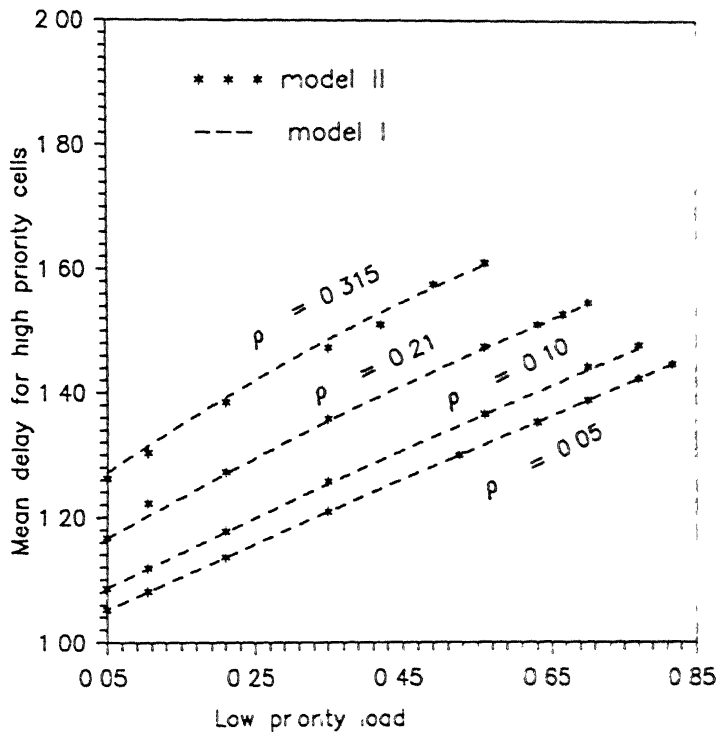


Fig 6.3 Mean delay for Q2 cells in an MMPP/D/1 priority system computed using models I(exact) and I(appr)

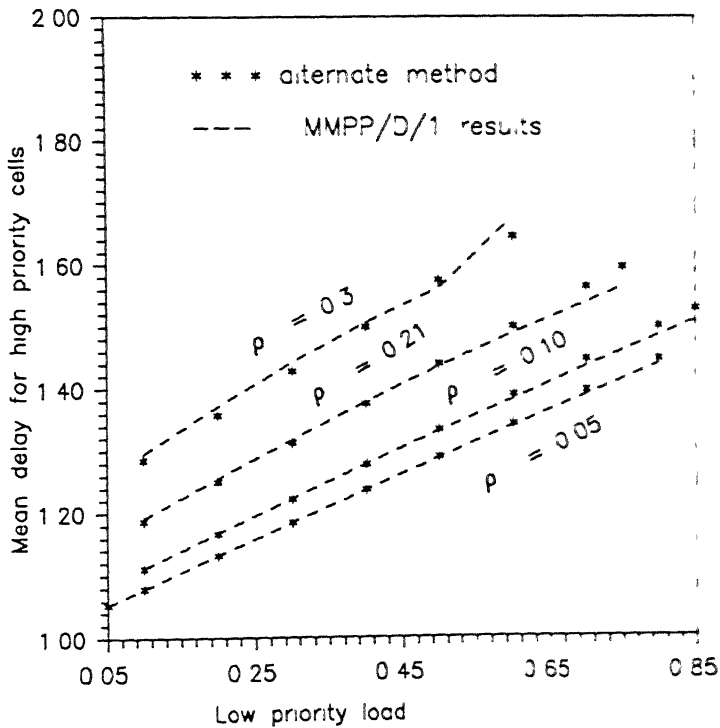


Fig 6.4 Mean delay for Q2 cells in an M/D/1 priority system computed using MMPP/D/1 results and alternate approach

CHAPTER 7

COMPUTATION OF THE QUEUE LENGTH DENSITIES OF A NON-PREEMPTIVE MMPP/D/1/K DUAL PRIORITY SYSTEM

7.1 INTRODUCTION

In this chapter, we study the characteristics of a non-preemptive MMPP/D/1/K dual priority system. The high and low priority customers are assumed to arrive at two separate queues Q2 and Q1 which have finite capacities of size N and M respectively. The service time per customer is assumed to be constant (D sec) for each class of customer. As in the infinite capacity case, the evaluation of the queue length densities at Q1 and Q2 is carried out using the matrix analytic approach. In the finite capacity case, the maximum number of customers with which the busy period of Q2 can start is limited to N (Q2 capacity). Hence the busy period distribution of the finite capacity system differs from that of the infinite capacity system. The equations developed in Chapter 3 are therefore modified to account for this. A recursive procedure for the evaluation of the busy period distribution (BPD) of the finite capacity queue is also developed. Finally, the BPD at Q2 and the QLDs at Q1 and Q2 are numerically computed for some typical examples and are compared with the results obtained using simulations.

As noted in Chapter 2, the non-preemptive MMPP/D/1 priority system has several similarities to the N/G/1 queue. By the same arguments it can be verified that the finite capacity priority system has several similarities to the N/G/1 finite capacity queue studied in detail in Blondia [1]. It differs from the N/G/1 finite capacity queue as follows. The evaluation of the QLD of Q1 requires the knowledge of the busy period distribution of Q2. The number of customers with which the busy period of Q2 starts depends on the traffic offered at Q1. The transition probability matrices pertaining to Q1 and Q2 are

coupled and the QLDs have to be evaluated iteratively

7.2 EMBEDDED SEMI- MARKOV SEQUENCE AND STATE TRANSITION PROBABILITY MATRICES OF Q1 AND Q2

Evaluation of the QLDs at Q1 and Q2 at the departure instants of the cells from their respective queues is carried out along the same lines as for the infinite capacity case. For brevity of notation, we consider the composite process obtained by superposition of the phase processes of the MMPPs to Q1 and Q2. As noted in Chapter 3, at any time t , given the phases of the MMPPs to Q1 and Q2, the phase of the composite process can be uniquely determined and vice versa. We define the various parameters pertaining to Q1 and Q2 in terms of the phase of the composite process. The symbols used to denote the various parameters of the finite capacity queueing system are the same as those used for the infinite capacity case. For ease of reference we summarize the parameters used. We shall assume the capacity of the queues Q1 and Q2 to be M and N respectively. It may be recalled that the case where Q1 capacity is finite has already been considered in Chapter 5. Hence, the modifications required to account for the finite size of Q2 are considered in more detail in this chapter. All parameters pertaining to Q1 are indicated by a superscript of (') and those of Q2 by ("). The notations used are -

(Q', Q'')	Infinitesimal generator matrices of MMPP1, MMPP2 to Q1 and Q2
Q^*	Infinitesimal generator matrix of the composite process
(Λ', Λ'')	Arrival rate matrices of MMPPs to Q1 and Q2
$(\underline{\Lambda}', \underline{\Lambda}'')$	Arrival rates at Q1 and Q2 corresponding to the different phases of the composite process
(τ'_n, τ''_n)	Departure epochs of cells from Q1 and Q2

(M, N)	No of phases of MMPP1, MMPP2
$J_n^{(1)'}, J_n^{(2)'}$	Phase of MMPP1, MMPP2 at τ_n'
$J_n^{(1)''}, J_n^{(2)'}$	Phase of MMPP1, MMPP2 at τ_n''
$\underline{J}(t)$	Phase of the composite process at time t
$\underline{J}_n', \underline{J}_n''$	Phase of the composite process at τ_n', τ_n''
$X'(t), X''(t)$	No of cells in Q1, Q2 at time t
X_n', X_n''	No of cells in Q1, Q2 at τ_n', τ_n''
$[P'(n,t)]_{1J}$	$P[n \text{ cells arrive at Q1 in } (0,t], \underline{J}(t)=J \mid \underline{J}(0)=1]$
$[P''(n,t)]_{1J}$	$P[n \text{ cells arrive at Q2 in } (0,t], \underline{J}(t)=J \mid \underline{J}(0)=1]$
$[x'']_{m1}$	$P[X_n'' = m, \underline{J}_n'' = 1]$
$[y_0'']_1$	$P[X''(t) = 0, \underline{J}(t) = 1] \text{ at any time } t$
p_0', p_0''	$P[Q1 \text{ empty at time } t], P[Q2 \text{ empty at time } t]$
$[U_k'(t)]_{1J}$	$P[\text{Busy period of Q1 starts and } k \text{ cells arrive at Q1 at or before time } t, \underline{J}(t) = J \mid \lambda'(0)=0, \underline{J}(0)=1]$
$[U_k''(t)]_{1J}$	$P[\text{Busy period of Q2 starts and } k \text{ cells arrive at Q2 at or before time } t, \underline{J}(t) = J \mid X''(0)=0, \underline{J}(0)=1]$
$[H'(t)]_{1J}$	$P[(\tau_n' - \tau_{n-1}') \leq t, \mid \underline{J}_{n-1}' = 1 \text{ and } X_{n-1}' > 0] \delta_{1J}$
$[H''(t)]_{1J}$	$P[(\tau_n'' - \tau_{n-1}'') \leq t, \mid \underline{J}_{n-1}'' = 1 \text{ and } X_{n-1}'' > 0] \delta_{1J}$
$u(t)$	unit step function
$\delta'(t), \delta_{1k}$	Dirac, Kronecker delta
I_{MN}, e	$MN \times MN$ identity matrix, $MN \times 1$ unit vector

The inter departure time of cells from Q1 $(\tau_n' - \tau_{n-1}')$ for $X_{n-1}' > 0$, consists of one Q1 cell service time and the busy period (BP) of Q2 if any. The duration of this BP depends on the phase of MMPP2 at τ_{n-1}' . Hence, the sequence $(X_n', J_n^{(1)'}, J_n^{(2)'})$ forms a SMC with the state space

$[0,1, M] \times [1,2, M] \times [1,2, N]$ Similarly $(X_n'', J_n^{(1)'}, J_n^{(2)'})$ forms an SMC with the state space $[0,1, N] \times [1,2, M] \times [1,2, N]$ In view of the uniqueness of the phase of the composite process, it can be shown that the sequences (X_n'', J_n'') and (X_n', J_n') also form SMCs with the state space $[0,1, N] \times [1,2, MN]$ and $[0,1, M] \times [1,2, MN]$ respectively The transition probability matrices of these SMCs are denoted as $Q''(t)$ and $Q'(t)$ respectively $Q''(t)$ can be expressed in terms of the $MN \times MN$ matrix mass functions $A_m''(t), B_m''(t)$ as follows

$$A_m''(t) = \int_{0-}^t dH''(\sigma) P''(m, \sigma) \quad m \geq 0, t \geq 0 \quad (7.2.1)$$

$$B_m''(t) = \sum_{k=1}^{m+1} U_k''(t-D) P''(m-k+1, D) u(t-D) \quad (7.2.2)$$

$$Q''(t) = \begin{bmatrix} B_0''(t) & B_1''(t) & B_2''(t) & B_{N-1}''(t) \sum_{m=N}^{\infty} B_m''(t) \\ A_0''(t) & A_1''(t) & A_2''(t) & A_{N-1}''(t) \sum_{m=N}^{\infty} A_m''(t) \\ 0 & A_0''(t) & A_1''(t) & A_{N-2}''(t) \sum_{m=N-1}^{\infty} A_m''(t) \\ 0 & 0 & A_0''(t) & A_{N-3}''(t) \sum_{m=N-2}^{\infty} A_m''(t) \\ 0 & 0 & 0 & A_0''(t) \sum_{m=1}^{\infty} A_m''(t) \end{bmatrix} \quad (7.2.3)$$

Similarly, $Q'(t)$ can be expressed in terms of the $MN \times MN$ matrix mass functions

$A_m'(t), B_m'(t)$ given by

$$A_m'(t) = \int_{0-}^t dH'(\sigma) P'(m, \sigma) \quad m \geq 0; t \geq 0 \quad (7.2.4)$$

$$B'_m(t) = \sum_{k=1}^{m+1} U'_k(t-D)P'(m-k+1,D)u(t-D) \quad (7.2.5)$$

In view of the constant service time requirement of D sec /customer, we get

$$H''(t) = u(t-D) I_{MN} \quad (7.2.6)$$

Computation of the elements of the $Q''(t)$ proceeds along the same lines as in Chapter 3 and are given by

$$A''_m(t) = P''(m,D)u(t-D) \quad (7.2.7)$$

$$B''_m(t) = \sum_{k=1}^{m+1} U''_k(t-D)P''(m-k+1,D)u(t-D) \quad (7.2.8)$$

$$\frac{d}{dt}U''_k(t) = \frac{(1-p'_0)}{D} \int_0^{Du(t-D_-)+tu(D-t)} P''(0,w)\underline{\Lambda}'' dw P''(k-1,t-w) + p'_0\delta_{1k}P''(0,t)\underline{\Lambda}'' \quad (7.2.9)$$

$$\underline{\Lambda}'' = I_M \otimes \Lambda'' \quad (7.2.10)$$

where I_M is the $M \times M$ identity matrix and \otimes is the Kronecker product

The corresponding equations for Q_1 are considered next. To determine the elements of $Q'(t)$ an expression for $H'(t)$ is required. Proceeding along the same lines as in Sec. 3.4 it can be verified that to determine $H''(t)$ an expression for the busy period distribution of Q_2 has to be developed first. The fact that Q_2 is of finite capacity makes the busy period distribution (BPD) of Q_2 to be different from that corresponding to the infinite capacity case. Since the inter departure time of cells from Q_1 depends upon the BPD of Q_2 , the elements of $Q'(t)$ are also different from that corresponding to the infinite capacity case.

Generalizing the $G(\cdot)$ matrices corresponding to the infinite capacity case, for the finite capacity Q_2 , we define $MN \times MN$ matrices $\underline{G}^{(N,k)}(t)$. The $(i,j)^{th}$ element of these matrices gives the probability that the BP of Q_2 is of duration at most t sec and ends with MMPP $\underline{2}$ in phase j given that the BP of Q_2 with a capacity of N started with k cells and with MMPP $\underline{2}$ in phase i . Since

the service time/customer is assumed to be constant (D sec), the busy period duration is an integral multiple of D . In view of this the distribution function of the busy period can be expressed as the weighted sum of the probabilities of the busy periods being of length D sec, $2D$ sec and so on. Towards this end we define $MN \times MN$ matrices $G_\ell^{(A,k)}$ for $\ell = 1, 2, 3, \dots$. The $(i, j)^{th}$ elements of these matrices denote the probability that the BP of Q2 of size is of duration ℓD sec and ends when the MMPP $\underline{2}$ is in phase j given that the BP started with k customers in Q2 with MMPP $\underline{2}$ in phase i . Using these

matrices $G_\ell^{(N,k)}(t)$, the distribution function of the BP of Q2 is given by-

$$G^{(N,k)}(t) = \sum_{\ell=k}^{\infty} G_\ell^{(N,k)} u(t-\ell D) \quad (7.2.11)$$

where $u(t)$ is the unit step function. In (7.2.11), the lower limit for ℓ is k as the busy period which starts with k customers has to be at least of duration kD seconds. As in Sec. 3.5, we define $MN \times 1$ vectors $c(k)$, the i^{th} elements of which are denoted as $c(k, i)$. Here $c(k, i)$ gives the joint probability that the BP of Q2 of capacity N is of length kD sec and starts with MMPP $\underline{2}$ in phase i . Using Fig. 3.3 and the arguments given in Sec. 3.5 it can be verified that $c(k, i)$ is given by

$$c(k, i) = \sum_{m=(1-\delta_{0k})}^k \sum_{j=1}^{MN} \left(\frac{1}{p_0''} y_0'' \right) \left(P_{j1}''(m, D) \right) \left(\sum_{\ell=1}^{MN} G_{k-1\ell}^{(A,m)} \right) \quad \text{for } k < N \quad (7.2.12)$$

$$\begin{aligned} &= \sum_{m=(1-\delta_{0k})}^{N-1} \sum_{j=1}^{MN} \left(\frac{1}{p_0''} y_0'' \right) \left(P_{j1}''(m, D) \right) \left(\sum_{\ell=1}^{MN} G_{k-1\ell}^{(N,m)} \right) \\ &\quad + \sum_{m=N}^{\infty} \sum_{j=1}^{MN} \left(\frac{1}{p_0''} y_0'' \right) \left(P_{j1}''(m, D) \right) \left(\sum_{\ell=1}^{MN} G_{k-1\ell}^{(N,N)} \right) \quad \text{for } k \geq N \quad (7.2.13) \end{aligned}$$

For ease of notation we define $G_k^{(N,0)}$ to be I , the $MN \times MN$ identity matrix. Let us compare the expression for $c(k, i)$ for both the infinite and finite capacity

cases Comparing (3.5.2) with (7.2.12) - (7.2.13) two differences can be noted Firstly, the matrix $G_\ell^{(m)}$ has been replaced by $G_\ell^{(\Lambda, m)}$ to account for the finite capacity of Q2 Secondly, $c(k, i)$ has two different expressions corresponding to the finite capacity case depending on the value of k The second term on the RHS of (7.2.13) is obtained as follows For $m > \Lambda$ (i.e. number of customers arriving at Q2 is greater than the capacity of Q2), only Λ of them receive service Next, let $C(k)$ denote the $MN \times MN$ diagonal matrices whose i^{th} diagonal element denotes the probability that BP of Q2 of capacity Λ is of duration kD sec given that it started at time τ'_n with MMPP $\underline{2}$ in phase ℓ Using (7.2.12) and (7.2.13), it can be verified that the $(i, j)^{\text{th}}$ element of $C(k)$ is given by

$$\begin{aligned} [C(k)]_{i,j} &= P[\text{BP of Q2 with capacity } \Lambda \text{ is of duration } kD \text{ sec} | \frac{j'}{n-1} = i] \delta_{i,j} \\ &= c(k, i) \left[\sum_{\ell=1}^{MN} c(k, \ell) \right]^{-1} \delta_{i,j} \end{aligned} \quad (7.2.14)$$

As noted earlier, given that the previous departure from Q1 left Q1 non-empty, the inter departure time of the next customer from Q1 being equal to t sec, implies that the intervening BP of Q2 is of duration $t-D$ sec In view of this $\frac{dH'(t)}{dt}$ is given by

$$\frac{dH'(t)}{dt} = C(k-1) \delta'(t-kD) \quad (7.2.15)$$

Using (7.2.15) in (7.2.4) we get

$$A'_n(t) = \sum_{k=1}^{\infty} u(t-kD) C(k-1) P'(n, kD) \quad (7.2.16)$$

An expression for $U'_k(t)$ can be obtained by considering the situation for the first cell arrival to Q1 when Q1 is empty but Q2 is not empty For this cell, the delay in service is equal to the sum of the residual service time of the Q2 cell currently being served and the additional BP (ABP) that the cells in Q2 may require after this (Note that service to the Q1 cell can start only

when Q2 is empty) As in Sec 3.5 we define $MN \times MN$ diagonal matrices $F(k)$ whose i^{th} diagonal element denotes the probability that ABP of Q2 of capacity N , is of length kD sec given that it started at time 0 with MMPP \underline{z} in phase 1. In the finite capacity case the maximum number of customers with which ABP can start is N . Hence the expression for $F(k)$ for the finite capacity case is different from that of the infinite case. Except for this difference, the expression for $U'_k(t)$ can be obtained by proceeding along the same lines as for the infinite capacity case (as in Sec 3.5) and we get -

$$\frac{dU'_k(t)}{dt} = \sum_{\iota=0}^{\infty} \sum_{n=(k-1)D}^{k-1} \left\{ \begin{aligned} & Du(t-\iota+1D_-) + (t-\iota D) u(\iota+1D-t) \\ & \left\{ u(t-\iota D) \int_0^{\infty} P'(0, t-\iota D-w) \underline{\Lambda}' dw \right. \\ & \left. P'(n, w) \right\} \end{aligned} \right\} \quad (7.2.17)$$

$$F(\iota) P'(k-n-1, \iota D) \left\{ \frac{(1-p_0'')}{D} + p_0'' \delta_{1k} P'(0, t) \underline{\Lambda}' \right\}$$

$$\underline{\Lambda}' = \Lambda' \otimes I_N \quad (7.2.18)$$

Finally, we obtain an expression for $F(\iota)$. Analogous to the computation of $C(k)$, we define $MN \times 1$ vectors $f(\iota)$ whose i^{th} element is denoted as $f(\iota, i)$ where $f(\iota, i)$ gives the probability that ABP of Q2 of capacity N , is of duration ιD sec and starts with MMPP \underline{z} in phase 1. It may be noted that the probability that the ABP starts with m customers and with the MMPP \underline{z} in phase 1 is given by x_{m1}'' . Given that the BP of Q2 of capacity N , starts with m customers and in phase 1, the probability that the BP is of duration ιD sec is given by the i^{th} row sum of the matrix $G_{\iota}^{(N, m)}$. Removing the condition on the number of customers with which the BP starts we get

$$f(\iota, i) = \sum_{m=1}^{\iota} \sum_{\ell=1}^{MN} x_{m1}'' G_{\iota \ell}^{(N, m)} \quad \text{for } \iota < N \quad (7.2.19)$$

$$= \sum_{m=1}^N \sum_{\ell=1}^{MN} x_{m1}'' G_{\ell_1 \ell}''^{(N,m)} \quad \text{for } i \geq N \quad (7.2.20)$$

The upper limit for m is N in (7.2.20) as the ABP cannot start with more than N customers for Q2 with capacity N . It may be recalled that the i^{th} diagonal element of $F(i)$ denotes the conditional probability where as the i^{th} element of $f(i)$ denotes the corresponding joint probability. Hence the diagonal elements of $F(i)$ are given by

$$[F(i)]_{ii} = f(i,1) \left[\sum_{\ell=1}^{MN} f(i,\ell) \right]^{-1} \quad (7.2.21)$$

7.3 CHARACTERISTICS OF THE BUSY PERIOD OF Q2

As in Chapter 4, the characteristics of the busy period of finite sized Q2 can be studied by considering the properties of the so called $G(\cdot)$ matrices. Generalizing the $G(\cdot)$ matrices corresponding to the infinite capacity case, for the finite capacity Q2, we define $MN \times MN$ matrices $G^{(N,k)}(\ell, t)$. The $(i, j)^{\text{th}}$ element of these matrices gives the probability that the BP of Q2 is of duration at most t sec, consists of ℓ services and ends with MMPP $\underline{2}$ in phase j given that the BP of Q2 with a capacity of N started with k cells and with MMPP $\underline{2}$ in phase i . The evaluation of the moment generating function of these matrices is considered first. Using this, the computation of the busy period distribution is considered next. Finally, the computation of the mean number of cells served during a busy period and the mean duration of the BP are considered.

PROBABILITY GENERATING FUNCTION OF $G^{(N,k)}(\ell, t)$

As in Chapter 4, we define the Markov renewal process $Q''(\cdot)$ to be in level i when the number of cells in Q2 is i and the phase of the MMPP $\underline{2}$ is j i.e. i is the set $\{1, j \mid 1 \leq j \leq MN\}$. We denote the LST of the z transform of

$$G^{(N,k)}_{(\ell,t)} \text{ as } \tilde{G}^{(N,k)}_{(z,s)} \text{ i.e.}$$

$$\tilde{G}^{(N,k)}_{(z,s)} = \sum_{\ell=1}^{\infty} \int_{t=0}^{\infty} e^{-st} dG^{(N,k)}_{(\ell,t)} \quad (7.3.1)$$

It is shown in Neuts [2] that for $N = \infty$, $\tilde{G}^{(\infty,k)}_{(z,s)}$ can be written as

$$\tilde{G}^{(\infty,k)}_{(z,s)} = \left[\tilde{G}^{(\infty,1)}_{(z,s)} \right]^k \quad (7.3.2)$$

This relationship is obtained by noting that for the infinite capacity case the first passage time from level 1 to 0 is identical to that from $i+1$ to i for $i > 0$. In the finite capacity case these two first passage times are not identical. This can be verified as follows. When the busy period of Q2 with capacity N starts in level 1 (i.e. with one customer in Q2), N customers can arrive during the service time of the first customer and all these customers will be accepted into the queue. Next let us consider the case when the BP of this queue starts in level i for $i > 1$. During the service time of the first customer of this busy period, only a maximum of $N + 1 - i$ customers are accepted into the queue. Hence the evolution of the busy period and the maximum number of new customers that can be accepted into the queue depends on the level of Q"(). Hence the first passage times are not identical. The evaluation of the double transform corresponding to the finite capacity case can be carried out exactly along the same lines as in Blondia [1]. This is because in the non-preemptive priority system considered here and in the N/G/1 finite capacity queue the busy period can in general start with more than one customer and the maximum number of customers with which the busy period can start is also finite.

Proceeding as in Blondia [1], next, we obtain a recursive procedure for the computation of the double transform of $G^{(N,k)}_{(\ell,t)}$. We first define $MN \times MN$ matrices denoted as $\mathcal{G}^{(n,k)}$, for $n=1,2,\dots,N$ and $k=1,2,\dots,N$. The $(i,j)^{\text{th}}$ elements of these matrices, denoted as $\mathcal{G}^{(n,k)}_{i,j}$, are random variables and they

denote the length of the busy period of Q2 which ends when MMPP $\underline{2}$ is in phase j given that the BP of Q2 with capacity n started with k cells in Q2 and with MMPP $\underline{2}$ in phase i . Let the distribution function of $\mathcal{G}^{(n,k)}$ be denoted as $\underline{G}^{(n,k)}(t)$. It may be noted that $\underline{G}^{(n,k)}(t)$ is equal to the marginal distribution function of $G^{(n,k)}(\ell, t)$ obtained by summing the later expression w r t ℓ i e

$$P[\mathcal{G}^{(n,k)} \leq t] = \underline{G}^{(n,k)}(t) = \sum_{\ell=1}^{\infty} G^{(n,k)}(\ell, t) \quad (7.3.3)$$

It may be noted that the $\mathcal{G}_{1,j}^{(n,k)}$ is the sum of k random variables. This can be verified as follows. Let us consider the computation of the time taken for $Q^{(n,k)}$ to go from level k to $k-1$ when Q2 capacity is n . During this first passage time only a maximum of $n-k+1$ new customers can be accepted into Q2. Hence the first passage from level k to $k-1$ with Q2 size of n is equal to the first passage time of Q2 of size $n-k+1$ from level 1 to 0 . By extending this observation it can be shown that the first passage time from level $k-1$ to $k-2$ with Q2 size of n is equal to the first passage time of Q2 of size $n-k+2$ from level 1 to 0 . By similar arguments it can be concluded that $\mathcal{G}_{1,j}^{(n,k)}$ is the sum of k random variables and writing the sum in matrix form we get

$$\mathcal{G}^{(n,k)} = \sum_{\ell=n-k+1}^n \mathcal{G}^{(\ell,1)} \quad (7.3.4)$$

Hence the generating function of $\mathcal{G}^{(n,k)}$ is given by the product of the generating functions of $\mathcal{G}^{(\ell,1)}$ for $\ell = n-k+1$ to n . In view of (7.3.4), and (7.3.3) we get

$$\tilde{G}^{(n,k)}(z,s) = \prod_{\ell=n-k+1}^n \tilde{G}^{(\ell,1)}(z,s) \quad (7.3.5)$$

Here (7.3.5) can be rewritten by writing out the first term of the product

and reusing (7.3.5) as follows

$$\tilde{G}^{(n,k)}_{(z,s)} = \tilde{G}^{(n-k+1,1)}_{(z,s)} \prod_{\ell=n-k+2}^n \tilde{G}^{(\ell,1)}_{(z,s)} \quad (7.3.6)$$

$$\tilde{G}^{(n,k)}_{(z,s)} = \tilde{G}^{(n-k+1,1)}_{(z,s)} \tilde{G}^{(n,k-1)}_{(z,s)} \quad (7.3.7)$$

Hence $\tilde{G}^{(n,k)}_{(z,s)}$ can be recursively computed using (7.3.7). Next, an expression for $\tilde{G}^{(n,1)}_{(z,s)}$ is obtained. Towards this end, we consider the various possible transitions from level 1 of Q2 of capacity n in a single service time. If no customers arrive during the service time, the BP ends after the service and the probability of this event is characterized by the matrix $\tilde{A}''_0(s)$. If k customers arrive during the first customer service and $k \leq n-1$, all these customers will be accepted into Q2 and will receive the service. When $k \geq n$, only n of these customers are accepted into Q2. The probability of these two events are characterized by the matrices $\tilde{A}''_k(s)$. Considering these three events we get

$$\tilde{G}^{(n,1)}_{(z,s)} = z \tilde{A}''_0(s) + z \sum_{k=1}^{n-1} \tilde{A}''_k(s) \tilde{G}^{(n,k)}_{(z,s)} + z \sum_{k=n}^{\infty} \tilde{A}''_k(s) \tilde{G}^{(n,n)}_{(z,s)} \quad (7.3.8)$$

Substituting (7.3.5) in (7.3.8) we get

$$\tilde{G}^{(n,1)}_{(z,s)} = z \left[\tilde{A}''_0(s) + \sum_{k=1}^{n-1} \tilde{A}''_k(s) \prod_{\ell=n-k+1}^n \tilde{G}^{(\ell,1)}_{(z,s)} + \sum_{k=n}^{\infty} \tilde{A}''_k(s) \prod_{\ell=1}^n \tilde{G}^{(\ell,1)}_{(z,s)} \right] \quad (7.3.9)$$

where (7.3.9) gives the functional equation satisfied by $\tilde{G}^{(n,1)}_{(z,s)}$

COMPUTATION OF THE BUSY PERIOD DISTRIBUTION OF Q2

As for the infinite capacity case, the busy period distribution can be computed by the approach given in Ramaswami [3]. Since this approach requires the computation of inverse LST, we shall consider an alternate approach

The method proposed for the infinite capacity case in Chapter 4 is modified here appropriately for computing the busy period distribution of the finite sized Q2. Since the service time/customer is assumed to be constant (D sec), the busy period duration is an integral multiple of D. In view of this the distribution function of the busy period can be expressed as the weighted sum of the probabilities of the busy periods being of length D sec, 2Dsec and so on. Towards this end we define $M \times M$ matrices $\underline{G}_\ell^{(N,k)}$ for $\ell = 1, 2, 3, \dots$. The $(i, j)^{th}$ elements of these matrices denote the probability that the BP of Q2 of size is of duration ℓD sec and ends when the MMPP $\underline{2}$ is in phase j given that the BP started with k customers in Q2 with MMPP $\underline{2}$ in phase i . Using these matrices $\underline{G}^{(N,k)}(t)$, the distribution function of the BP of Q2 is given by

$$\underline{G}^{(N,k)}(t) = \sum_{\ell=k}^{\infty} \underline{G}_\ell^{(N,k)} u(t-\ell D) \quad (7.3.10)$$

where $u(t)$ is the unit step function. In (7.3.10), the lower limit for ℓ is k as the busy period which starts with k customers has to be at least of duration kD sec. In view of (7.3.10), finding the busy period distribution of Q2 reduces to the computation of the matrices $\underline{G}_\ell^{(N,k)}$ for $\ell = 1, 2, 3, \dots$. Using (7.3.1), (7.3.3) and (7.3.10) it can be verified that the LST of $\underline{G}^{(N,k)}(t)$, denoted as $\tilde{\underline{G}}^{(N,k)}(s)$, is given by

$$\tilde{\underline{G}}^{(N,k)}(s) = \tilde{\underline{G}}^{(N,k)}(1, s) \quad (7.3.11)$$

$$= \sum_{\ell=k}^{\infty} \underline{G}_\ell^{(N,k)} e^{-s\ell D} \quad (7.3.12)$$

Evaluating $\tilde{\underline{G}}^{(n,1)}(z, s)$ at $z=1$ and using (7.3.11)-(7.3.12) and (3.4.2) in (7.3.8) we get

$$\sum_{\ell=1}^{\infty} \underline{G}_\ell^{(n,1)} e^{-s\ell D} = P''(0, D) e^{-sD} + \sum_{k=1}^{n-1} P''(k, D) e^{-sD} \sum_{\ell=k}^{\infty} \underline{G}_\ell^{(n,k)} e^{-s\ell D}$$

$$+ \sum_{k=n}^{\infty} P''(k,D)e^{-sD} \sum_{\ell=1}^{\infty} G_{\ell}^{(n,n)} e^{-s\ell D} \quad (7.3.13)$$

Comparing the coefficient of $[e^{-sD}]^m$ on both sides of (7.3.13) we get

$$G_0^{(n,1)} = P''(0,D) \quad \text{for } m=1 \quad (7.3.14)$$

$$G_m^{(n,1)} = \sum_{k=1}^{n-1} P''(k,D) G_{m-1}^{(n,k)} + \sum_{k=n}^{\infty} P''(k,D) e^{-sD} G_{m-1}^{(n,n)} \quad \text{for } m>1 \quad (7.3.15)$$

A procedure for the evaluation of the matrices $G_m^{(n,1)}$ using (7.3.14)-(7.3.15) can be formulated better by defining two sets of infinite dimensional matrix arrays $\{Y(1), Y(2), \dots, Y(n)\}$ and $\{X(1), X(2), \dots, X(n)\}$. The matrix arrays $Y(i)$'s give the busy period distribution of Q2 starting with a single customer when the Q2 capacity is i . In other words the j^{th} element of the array $Y(i)$ viz $Y(i,j)$ denotes the probability that the duration of the BP of Q2 of capacity i starting with a single customer is jD sec i.e. $Y(i,j)$ is given by

$$Y(i,j) = G_j^{(i,1)} \quad (7.3.16)$$

$Y(i,j)$ is a matrix as the phase of the MMPP \underline{Z} must also be kept track of at the beginning and the end of the BP. The matrix arrays $X(i)$'s give the busy period distribution of Q2 of capacity n when the number of customers at the beginning of the BP of Q2 is i . In other words the j^{th} element of the array $X(i)$ viz $X(i,j)$ denotes the probability that the duration of the BP of Q2 of capacity n starting with i customers is jD sec i.e. $X(i,j)$ is given by

$$X(i,j) = G_j^{(n,i)} \quad (7.3.17)$$

Using (7.2.16) and (7.3.17) it may be verified that

$$Y(n) = X(1) \quad (7.3.18)$$

Substituting (7.3.16)-(7.3.18) in (7.3.14)-(7.3.15) we get

$$Y(n,1) = P''(0,D) \quad (7.3.19)$$

$$Y(n,m) = \sum_{k=1}^{n-1} P''(k,D)X(k,m-1) + \sum_{k=n}^{\infty} P''(k,D)X(n,m-1) \quad \text{for } m > 1 \quad (7.3.20)$$

To compute $Y(n,m)$ using (7.3.20), $X(k,m-1)$ for $k = 1, 2, \dots, n$ should be known. These matrices can be found by evaluating (7.3.7) at $z=1$. Substituting (7.3.11) in (7.3.7), evaluated at $z=1$, we get

$$\tilde{G}^{(n,k)}(s) = \tilde{G}^{(n-k+1,1)}(s) \tilde{G}^{(n,k-1)}(s) \quad (7.3.21)$$

Substituting (7.3.8) in (7.3.21) we get

$$\sum_{\ell=k}^{\infty} G_{\ell}^{(n,k)} e^{-s\ell D} = \sum_{\ell=1}^{\infty} G_{\ell}^{(n-k+1,1)} e^{-s\ell D} \sum_{\ell'=k-1}^{\infty} G_{\ell'}^{(n,k-1)} e^{-s\ell' D} \quad (7.3.22)$$

Comparing the coefficient of e^{-smD} on both sides of (7.3.22) we get

$$G_m^{(n,k)} = \sum_{\ell=1}^{m-1} G_{\ell}^{(n-k+1,1)} G_{m-\ell}^{(n,k-1)} \quad (7.3.23)$$

Using (7.3.16) and (7.3.17) in (7.3.23) we get

$$X(k,m) = \sum_{\ell=1}^{m-1} Y(n-k+1,\ell)X(k-1,m-\ell) \quad \text{for } m \geq k \quad (7.3.24)$$

$$= 0 \quad \text{for } m < k \quad (7.3.25)$$

Here (7.3.25) is obtained by noting that a BP which starts with k customers should be at least of duration kD sec. It may be noted that, for $k > 2$, $X(k,m)$ can be recursively computed using (7.3.24) and (7.3.25) if the values of the first $m-1$ elements of the matrix arrays $Y(1)$, $Y(2)$, \dots , $Y(n-k+1)$ are known. It may be recalled that $X(1,m-1)$ is equal to $Y(n,m-1)$. Noting that the maximum value of k is n (Q2 size), it can be concluded that the m th element of the arrays $X(k)$ for $k = 1, 2, \dots, n$ can be determined if the first $(m-1)$ elements of

the arrays $Y(k)$, for $k = 1, 2, \dots, n$ are known

It may be recalled that we observed earlier that to compute $Y(n, m)$ using (7.3.20), $X(k, m-1)$ for $k = 1, 2, \dots, n$ should be known. This implies that $Y(n, m)$ can be recursively computed and the recursive procedure can be stated as follows -

- (1) Let $n = 1$
- (2) Compute $Y(1, 1)$ using (7.3.19). Set $X(k, 1)$ to be zero for $k > 1$
- (3) Compute $Y(1, k)$ for $k = 2, 3, \dots$ using (7.3.20) until $Y(1, k)$ is less than a threshold η
- (4) Increment n . Compute $Y(n, 1)$ using (7.3.19). Set $X(k, 1)$ to be zero for $k > 1$. Let $\ell = 1$
- (5) Increment ℓ and Compute $Y(n, \ell)$ using (7.3.20) and $X(k, \ell)$ using (7.3.24)-(7.3.25)
- (6) If $Y(n, \ell) > \eta$ goto step 5. Else go to step 4 if $n < N$ (the actual buffer size of Q2)

It may be noted that for computing the busy period distribution of Q2 of size N , the busy period distribution of Q2 of size $1, 2, \dots, N-1$ should first be found. The recursive procedure for the computation of $Y(n)$ also yields the matrix arrays $X(k)$ for $k = 1, 2, 3, 4, \dots, N$. Hence the busy period distribution of Q2 of size N starting with k customers is also computed as a by product of the recursive procedure.

Next we consider a technique for minimizing the computational complexity required for implementing the recursive procedure given above. We shall refer to this method as the indirect method and the method which uses the recursive procedure directly is referred to as direct method. It may be recalled that the $(i, j)^{\text{th}}$ element of $G_{\ell}^{(n, k)}$ is given by

$$\begin{aligned} \left[\hat{G}_\ell^{(n,k)} \right]_{1,j} &= P[\text{BP of Q2 of capacity } n \text{ is of length } \ell D \text{ sec} \\ &\quad \text{and ends with MMPP 2 in phase } j \mid \text{it started} \\ &\quad \text{with } k \text{ customers in phase } i] \end{aligned} \quad (7.3.26)$$

In view of the one-one mapping of the phases of MMPP 2 to those of MMPP 1 and MMPP 2, (7.3.26) can be rewritten as

$$\begin{aligned} \left[\hat{G}_\ell^{(n,1)} \right]_{1,j} &= P[\text{BP of Q2 of capacity } n \text{ is of length } \ell D \text{ sec and ends with} \\ &\quad \text{MMPP 2 in phase } j'' \text{ and MMPP 1 in phase } j' \mid \text{it started with} \\ &\quad k \text{ customers and MMPP 2 in phase } i'' \text{ and MMPP 1 in phase } i'] \end{aligned} \quad (7.3.27)$$

$$\begin{aligned} &= \{P[\text{BP of Q2 of capacity } n \text{ is of length } \ell D \text{ sec and ends with} \\ &\quad \text{MMPP 2 in phase } j'' \mid \text{it started with } k \text{ customers in phase } i'']\} \\ &\quad \{P[\text{The phase of MMPP 1 is } j' \text{ at time } \ell D \mid \text{it is} \\ &\quad \text{in phase } i' \text{ at time } 0]\} \end{aligned} \quad (7.3.28)$$

where $i = (i'-1)M + i''$ and $j = (j'-1)M + j''$ (7.3.28) is obtained from

(7.3.27) by noting that the phases of MMPP 1 and MMPP 2 are independent

Let us define $N \times N$ matrices $\hat{G}_\ell^{(n,k)}$ whose (i,j) th elements are given by

$$\begin{aligned} \left[\hat{G}_\ell^{(n,k)} \right]_{1,j} &= P[\text{BP of Q2 of capacity } n \text{ is of length } \ell D \text{ sec} \\ &\quad \text{and ends with MMPP 2 in phase } j \mid \text{it started} \\ &\quad \text{with } k \text{ customers in phase } i] \end{aligned} \quad (7.3.29)$$

It may be noted that the recursive procedure that we have arrived at for the computation of the busy period distribution is independent of the number of phases of the MMPP. Hence the above recursive procedure can also be used for

the computation of $\hat{G}_\ell^{(n,k)}$. Using (3.6.2) it may be noted that

$$\begin{aligned} \left[e^{Q^* \ell D} \right]_{1,j} &= \{P[\text{The phase of MMPP 1 is } j \text{ at time } \ell D \mid \text{it is} \\ &\quad \text{in phase } i \text{ at time } 0]\} \end{aligned} \quad (7.3.30)$$

Using (7.3.29) and (7.3.30) in (7.3.28) we get

$$\left[G_{\ell}^{(n,k)} \right]_{i,j} = \left[\hat{G}_{\ell}^{(n,k)} \right]_{i',j'} \left[e^{Q^{**}LD} \right]_{i'',j''} \quad (7.3.31)$$

By computing (7.3.31) for various values of i and j it can be verified that (7.3.31) can be written in matrix form as

$$G_{\ell}^{(n,k)} = e^{Q^{**}LD} \otimes \hat{G}_{\ell}^{(n,k)} \quad (7.3.32)$$

where \otimes is the kronecker product of matrices. Hence the computation of the $MN \times MN$ matrices $G_{\ell}^{(n,k)}$ is reduced to the computation of $N \times N$ matrices $\hat{G}_{\ell}^{(n,k)}$. As noted earlier the computation of the busy period distribution of Q_2 of capacity n requires the computation of the BPD of Q_2 of capacities 1 to $n-1$ as a prerequisite. Hence computational savings of the order of $O(M^2)$ results in each of these n stages of computation. It may be recalled that M denotes the total number of phases of MMPP-1. The computation of the matrices $e^{Q^{**}LD}$ is considered in Sec. 5.8. The computation of these matrices and performing the kronecker product of (7.3.32) are the overhead involved in the indirect method. This overhead is expected to be small compared to the savings in the computation referred to above. It should be pointed out here that even for the infinite capacity case the indirect method is applicable. However, since the recursive procedure is run only once, the reduction in the computational effort with the indirect method is not worth the complexity in the implementation and hence was not considered earlier.

COMPUTATION OF THE AVERAGE DURATION AND THE NUMBER OF CUSTOMERS SERVED DURING THE BUSY PERIODS OF Q_2

The computation of the average duration and the number of customers served during the busy periods of a finite capacity Q_2 can be carried out along the same lines as for the infinite capacity case considered in Sec. 4.4. We define the $MN \times MN$ matrices of mass function $L''(k,t)$ whose $(j,j')^{\text{th}}$ element

gives the probability that the first busy period of Q2 of capacity N is of duration at most t sec and consists of k services and ends with MMPP $\underline{2}$ in phase j' given that it started in phase j . We denote the double transform of $L^*(k,t)$ as $\tilde{L}^*(z,s)$. Considering all the possible ways in which the BP of Q2 can start, we get

$$\tilde{L}^*(z,s) = \sum_{k=1}^N \tilde{U}_k^*(0) \tilde{G}^{(N,k)}(z,s) + \sum_{k=N+1}^{\infty} \tilde{U}_k^*(0) \tilde{G}^{(N,N)}(z,s) \quad (7.3.34)$$

The second term on the RHS of (7.3.34) is obtained by noting that when more than N customers arrive before the busy period starts only N of them are allowed into Q2. Let $\tilde{\mu}_1^*$ and μ_1^* be the $M \times 1$ vectors whose i^{th} elements denote respectively, the mean number served during and the mean duration of the first busy period of Q2 given that $X^*(0) = 0$ and $\underline{1}^*(0) = 1$. $\tilde{\mu}_1^*$ can be obtained by differentiating (7.3.34) with respect to z . It may be noted that a busy period of Q2 of capacity n starting with k customers will eventually end with probability one. Therefore $\tilde{G}^{(n,k)}(1,0)$ is stochastic and hence we get

$$\tilde{G}^{(n,k)}(1,0)e = e \quad 1 \leq n \leq N, \quad 1 \leq k \leq n \quad (7.3.35)$$

Differentiating (7.3.34) w.r.t z and using (7.3.5)-(7.3.7) we get

$$\begin{aligned} \left. \frac{\partial}{\partial z} \tilde{L}^*(z,s)e \right|_{z=1,s=0} &= \tilde{\mu}_1^* = \sum_{k=1}^N \tilde{U}_k^*(0) \left[\frac{\partial}{\partial z} \tilde{G}^{(N-k+1,1)}(z,s) \tilde{G}^{(N,k-1)}(1,0)e \right. \\ &\quad + \tilde{G}^{(N-k+1,1)}(1,0) \frac{\partial}{\partial z} \tilde{G}^{(N-k+2,1)}(z,s) \tilde{G}^{(N,k-2)}(1,0)e + \\ &\quad \left. + \prod_{\ell=N-k+1}^{N-1} \tilde{G}^{(\ell,1)}(1,0) \frac{\partial}{\partial z} \tilde{G}^{(N,1)}(z,s)e \right] \Bigg|_{z=1,s=0} \\ &\quad + \sum_{k=N+1}^{\infty} \tilde{U}_k^*(0) \frac{\partial}{\partial z} \tilde{G}^{(N,N)}(z,s)e \Bigg|_{z=1,s=0} \end{aligned} \quad (7.3.36)$$

Let

$$\tilde{G}^{(n,k)} = \tilde{G}^{(n,k)}_{(1,0)} \quad (7.3.37)$$

$$\overline{\tilde{\mu}^n} = \left. \frac{\delta}{\delta z} \tilde{G}^{(n,1)}_{(z,s)} e \right|_{z=1, s=0} \quad (7.3.38)$$

For notational convenience, let $\tilde{G}^{(n,0)} = I$. Using (7.3.35), (7.3.37)-(7.3.38) and combining the like terms we get

$$\tilde{\mu}_1^n = \sum_{k=1}^N \tilde{U}_k^{(0)} \left\{ \sum_{j=0}^{k-1} \prod_{i=N-k+1}^{N-k+j} \tilde{G}^{(i,1)} \overline{\tilde{\mu}^{N-k+j+1}} \right\} + \sum_{k=N+1}^{\infty} \tilde{U}_k^{(0)} \left\{ \sum_{j=0}^{N-1} \prod_{i=1}^j \tilde{G}^{(i,1)} \overline{\tilde{\mu}^{j+1}} \right\} \quad (7.3.39)$$

The term in the second braces of (7.3.39) is obtained by substituting $k=N$ in the previous term inside the braces. This follows by comparing the first and the second terms of (7.3.34). Next, using (7.3.5) in (7.3.39) we get

$$\tilde{\mu}_1^n = \sum_{k=1}^N \tilde{U}_k^{(0)} \sum_{j=0}^{k-1} \tilde{G}^{(N-k+j,j)} \overline{\tilde{\mu}^{N-k+j+1}} + \sum_{k=N+1}^{\infty} \tilde{U}_k^{(0)} \sum_{j=0}^{N-1} \tilde{G}^{(j,j)} \overline{\tilde{\mu}^{j+1}} \quad (7.3.40)$$

From (7.3.40) it can be noted that for the computation of $\tilde{\mu}^n$, the parameters $\tilde{G}^{(n,k)}$ and $\overline{\tilde{\mu}^n}$ have to be evaluated and these are considered next.

First, the evaluation of $\tilde{G}^{(n,k)}$ is considered. It may be noted from (7.3.6) that $\tilde{G}^{(m,k)}$, for $m=1,2,\dots,N$ and $k=1,2,\dots,m$, can be recursively computed if $\tilde{G}^{(n,1)}$ for $n=1,2,\dots,m$ is known. $\tilde{G}^{(n,1)}$ can be evaluated as follows.

Let $n > 1$. Evaluating (7.3.9) at $z=1, s=0$ and using (7.3.37) we get

$$\tilde{G}^{(n,1)} = \tilde{A}_0^n + \sum_{k=1}^{n-1} \tilde{A}_k^n \prod_{\ell=n-k+1}^n \tilde{G}^{(\ell,1)} + \sum_{k=n}^{\infty} \tilde{A}_k^n \prod_{\ell=1}^n \tilde{G}^{(\ell,1)} \quad (7.3.41)$$

$$= \tilde{A}_0^n + \tilde{A}_1^n \tilde{G}^{(n,1)} + \sum_{k=2}^{n-1} \tilde{A}_k^n \prod_{\ell=n-k+1}^{n-1} \tilde{G}^{(\ell,1)} \tilde{G}^{(n,1)} + \sum_{k=n}^{\infty} \tilde{A}_k^n \prod_{\ell=1}^{n-1} \tilde{G}^{(\ell,1)} \tilde{G}^{(n,1)} \quad (7.3.42)$$

where $\tilde{\mathbf{A}}_k'' = \tilde{\mathbf{A}}_k''(0)$ (7.3.42) is obtained from (7.3.41) by taking out the term corresponding to $k=1$ in the summation and removing the term of $\ell=n$ in the product. Combining the like terms we get

$$\tilde{\mathbf{G}}^{(n,1)} = \left[\mathbf{I} - \tilde{\mathbf{A}}_1'' - \sum_{k=2}^{n-1} \tilde{\mathbf{A}}_k'' \prod_{\ell=n-k+1}^{n-1} \tilde{\mathbf{G}}^{(\ell,1)} - \sum_{k=n}^{\infty} \tilde{\mathbf{A}}_k'' \prod_{\ell=1}^{n-1} \tilde{\mathbf{G}}^{(\ell,1)} \right]^{-1} \tilde{\mathbf{A}}_0'' \quad \text{for } n > 1 \quad (7.3.43)$$

When $n=1$, the second term of (7.3.41) is zero and in this case we get

$$\tilde{\mathbf{G}}^{(1,1)} = \left[\mathbf{I} - \sum_{k=n}^{\infty} \tilde{\mathbf{A}}_k'' \right]^{-1} \tilde{\mathbf{A}}_0'' \quad (7.3.44)$$

The inverses of (7.3.43) and (7.3.44) exist as they have the form $[\mathbf{I} - \mathfrak{F}]^{-1}$, where \mathfrak{F} is a sub-stochastic matrix. Finally, it may be noted that $\tilde{\mathbf{G}}^{(n,1)}$ can be recursively computed using (7.3.44) and (7.3.43).

Next, the computation of $\overline{\tilde{\mu}^n}$ is considered. From (7.3.38) it can be noted that the 1th element of this vector gives the average number of customers served during a busy period of Q2 of capacity n which starts with the MMPP $\underline{2}$ in phase 1. Differentiating (7.3.9) w.r.t z , setting $z=1, s=0$ and postmultiplying by \mathbf{e} , $\overline{\tilde{\mu}^n}$ is given by

$$\overline{\tilde{\mu}^n} = \sum_{k=0}^{\infty} \tilde{\mathbf{A}}_k'' \mathbf{e} + \sum_{k=1}^{n-1} \tilde{\mathbf{A}}_k'' \sum_{j=0}^{k-1} \prod_{i=n-k+1}^{n-1} \tilde{\mathbf{G}}^{(i,1)} \overline{\tilde{\mu}^{n-k+j+1}} + \sum_{k=n}^{\infty} \tilde{\mathbf{A}}_k'' \sum_{j=0}^{n-1} \prod_{i=1}^j \tilde{\mathbf{G}}^{(i,1)} \overline{\tilde{\mu}^{j+1}} \quad (7.3.45)$$

To simplify this further, the term corresponding to $k=1$ is written out separately. Next, in the second and third terms the term corresponding to $j=k-1$ and $n-1$ are separated out. With this (7.3.45) becomes

$$\overline{\tilde{\mu}^n} = \sum_{k=0}^{\infty} \tilde{\mathbf{A}}_k'' \mathbf{e} + \tilde{\mathbf{A}}_1'' \overline{\tilde{\mu}^n} + \sum_{k=2}^{n-1} \tilde{\mathbf{A}}_k'' \prod_{i=n-k+1}^{n-1} \tilde{\mathbf{G}}^{(i,1)} \overline{\tilde{\mu}^n} + \sum_{k=n}^{\infty} \tilde{\mathbf{A}}_k'' \prod_{i=1}^{n-1} \tilde{\mathbf{G}}^{(i,1)} \overline{\tilde{\mu}^n}$$

$$+ \sum_{k=2}^{n-1} \tilde{A}_k^n \sum_{j=0}^{k-2} \prod_{i=n-k+1}^{n-k+j} \tilde{G}^{(i,1)} \overline{\tilde{\mu}^{n-k+j+1}} + \sum_{k=n}^{\infty} \tilde{A}_k^n \sum_{j=0}^{n-2} \prod_{i=1}^{n-2-j} \tilde{G}^{(1,1)} \overline{\tilde{\mu}^{j+1}} \quad (7.3.46)$$

Using (7.3.43)-(7.3.44) and (7.3.5) in (7.3.46) we get

$$\begin{aligned} \overline{\tilde{\mu}^n} = & \tilde{G}^{(n,1)} \left[\tilde{A}_0^n \right]^{-1} \left\{ e + \sum_{k=2}^{n-1} \tilde{A}_k^n \sum_{j=0}^{k-2} \tilde{G}^{(n-k+j,j)} \overline{\tilde{\mu}^{n-k+j+1}} \right. \\ & \left. + \sum_{k=n}^{\infty} \tilde{A}_k^n \sum_{j=0}^{n-2} \tilde{G}^{(j,j)} \overline{\tilde{\mu}^{j+1}} \right\} \quad (7.3.47) \end{aligned}$$

$\overline{\tilde{\mu}^n}$ can be recursively computed using (7.3.47). This completes the computation of $\tilde{\mu}_1^n$. The average duration of the busy period can also be computed along the same lines. It should be pointed out here that our approach for the computation of the average number served during the busy period follows closely the approach given in Blondia [1] for the N/G/1 finite capacity queue.

7.4. COMPUTATION OF THE STATIONARY QUEUE LENGTH DENSITIES AT Q1 AND Q2

Computation of the vectors x_k^n and x_k' as well as y_0^n are considered first in this section. As in Chapter 5, an approximate model is considered for the finite capacity case as well. The numerical results obtained through numerical computation as well as through simulations are presented next. Knowing the values of y_0' and y_0^n the average queueing delays at Q1 and Q2 are computed using the results of Chapter 5. It may be recalled that the i^{th} element of x_k^n gives the probability of finding the queue length at Q2 to be k and the phase of MMPP $\underline{2}$ to be i at a departure instant of Q2. The vector x_k' is similarly defined.

COMPUTATION OF x_0'' AND y_0''

The computation of x_0'' is carried out along the same lines as for the infinite capacity case considered in Sec 5.2. We define $MN \times MN$ matrix mass functions $K_0''(n, t)$ whose (j, j') th element gives the conditional probability, given that the busy cycle of Q2 starts with MMPP $\underline{2}$ in phase j , that the busy cycle consists of n services, is of duration at most t and ends in phase j' . Let $\tilde{K}_0''(z, s)$ denote the double transform of $K_0''(n, t)$. Proceeding along the same lines as in Sec 4.6 it can be shown that

$$\tilde{K}_0''(z, s) = \sum_{k=1}^N \tilde{U}_k''(s) \tilde{G}^{(N, k)}(z, s) + \sum_{k=N+1}^{\infty} \tilde{U}_k''(s) \tilde{G}^{(N, N)}(z, s) \quad (7.4.1)$$

The probability $x''(0, j)$, the j th element of x_0'' , is the inverse of the mean recurrence time the state $(0, j)$ of the Markov chain $Q''(\infty)$. Whenever $Q''(\infty)$ visits the state $(0, j)$, the Markov renewal process of the lattice type $\tilde{K}''(z, 0)$ also visits the state $(0, j)$. Hence the MRT of the state $(0, j)$ of $Q''(\infty)$ and $\tilde{K}''(z, 0)$ are the same. It may be noted that $\tilde{K}_0''(z, 0)$ is equal to $\tilde{L}''(z, 0)$ by comparing (7.3.34) and (7.4.1). By applying Theorem 2.11 of Hunter [4], the mean recurrence time of $(0, j)$, denoted as $m''(0, j)$, is given by

$$m''(0, j) = (k_0'' \tilde{\mu}_1'') [(k_0'')_j]^{-1} \quad (7.4.2)$$

where $\tilde{\mu}_1''$ is given by (7.3.40) and $(k_0'')_j$ denotes the j th element of k_0'' . k_0'' is the invariant probability vector of $\tilde{K}_0''(1, 0)$. Hence x_0'' is given by

$$x_0'' = \left[k_0'' \tilde{\mu}_1'' \right]^{-1} k_0'' \quad (7.4.3)$$

The computation of y_0'' proceeds along the same lines as in Chapter 5. In the present case the parameters corresponding to the finite capacity case has to be used. Since the arguments are essentially the same, for the finite capacity case as well, y_0'' is given by (5.5.9). The expressions for x_0' and y_0'

corresponding to both finite and infinite capacity case are considered in Chapter 5 and hence they are not reproduced here

COMPUTATION OF THE QLDs USING THE APPROXIMATE MODEL

As in Chapter 5, we consider an approximate model for the computation of the QLDs of Q1 and Q2, when Q2 has finite capacity. As in the infinite capacity case, we treat the inter departure time of cells from a non-empty Q1 to consist of busy periods of a hypothetical process. The busy period distribution of the hypothetical process is obtained as the weighted average of the BPDs of Q2 of capacity N , corresponding to each possible initial phase of MMPP 2. In this case the matrices $A_m''(t)$, $B_m''(t)$ and $P''(m,t)$ become $N \times N$ matrices and depend only on the phase of MMPP 2 but not that of MMPP 1. Similarly, $A_m'(t)$, $B_m'(t)$ and $P'(m,t)$ become $M \times M$ matrices and depend only on the phase of MMPP 1 but not that of MMPP 2. The recursive procedure for the computation of BPD of Q2 of capacity N is also valid for this case. In this case the parameters of the approximate model should be used in the equations given in Sec 7.3 and the phase of MMPP 2 should be replaced by that of MMPP 2.

NUMERICAL RESULTS

The results on the computation of the busy period distribution of Q2 and the queue length densities at Q1 and Q2 as well as their average delays are presented in this section. The results obtained through the numerical computations are compared with those obtained using simulation. We refer to the exact model and the approximate model as model I and II respectively. The traffic to Q1 and Q2 are assumed to originate from N_1 , N_2 on/off sources. We assume the on/off sources to be of identical type. However, the equations given in this chapter are also valid for the case where the on/off sources to

Q1 and Q2 are dissimilar. The on/off sources have average on duration, % on duration and bit rate during on duration as 33 msec, 35% and 1 Mbps respectively. The on and off durations are exponentially distributed. A cell size of 53 bytes and output link capacity of 150 Mbps are also assumed. The composite traffic at Q1 and Q2 are approximated by two independent 2 phase MMPPs using the method proposed in Heffes [5].

First, the evaluation of the busy period distribution of Q2 of capacity λ starting with one customer is considered. It may be noted that when the busy period starts with a single customer, the busy period distribution of Q2 is the same as that of the MMPP/D/1/K queue with the FCFS discipline. The recursive procedure of Sec 7.3 gives only the conditional BPD i.e the BPD given that the busy period starts in phase i for $i=1,2, \dots, MN$. In order to obtain the unconditional BPD we should know the probability of the BP starting in phase i . Since, Q1 and Q2 are coupled, this information cannot be obtained without solving for the QLDs at Q1 and Q2. However, for the FCFS queue, this probability can be readily obtained. If we plot the conditional BPD the need for the FCFS queue does not arise. However, in order to present the results conveniently, the busy period distribution of the equivalent FCFS queue is considered.

Let N be chosen to be 25. Let N_1 and N_2 be 180, 180. Corresponding to this case, the parameters of the 2 phase MMPPs (λ', Q'') and (λ'', Q''') are found. Using (3.1.1), (3.1.5) and (3.1.6), the parameters of MMPP 1 and MMPP 2 are found. The parameters of the MMPP to the FCFS queue is chosen to be the same as that of MMPP 2. Using the recursive procedure given in Sec 7.3, the conditional busy period distribution at Q2 is found using both the direct and indirect method. The unconditional BPD is found using the equation given by-

$$P(\text{BP}=\text{nD}) = \frac{y_0 G_n^{(N,1)}}{y_0 e} \quad (7.4.4)$$

where the i^{th} element of y_0 gives the probability that the FCFS queue is empty and the MMPP is in phase i at an arbitrary time instant. The results from both the methods are found to match well and the indirect method is found to require about 50% less computation time than the direct method. The probability mass function (PMF) of the unconditional busy period computed above is shown in Fig. 7.1. As both the indirect and direct methods agree we have shown only one curve corresponding to the computations. The busy period is also computed using the simulation routine discussed in Sec. 5.9. It may be noted that the computation of the BPD using the simulation routine does not require any additional effort. In the infinite capacity case the capacity of Q2 is chosen to be an arbitrarily large value. In the finite capacity case it is chosen to be N . The cells which arrive when Q2 is full are discarded. The results obtained using simulation corresponding to $N_2=180$ is also shown in Fig. 7.1. The results obtained using both computation and simulation corresponding to $N_1=150$ and $N_2 = 240$ are shown in Fig. 7.2. (It may be noted that for the simulation routine the value of N_1 is not required, for computing the parameters of MMPP $\underline{2}$ it is required). From these two figures it can be concluded that the results obtained through the recursive procedure match well with the simulation results.

Next, the results on the evaluation of the QLDs at Q1 and Q2 are considered. We first consider two examples in which the capacities of Q1 and Q2 are ∞ and 25 respectively. In the first example, N_1 and N_2 are chosen to be both equal to 180. The traffic offered to both Q1 and Q2 are equal to 0.42. The MMPP model parameters are found as mentioned above. The QLDs are computed using the equations given in Sec. 7.2 - 7.4. For the computation of the

invariant probability vectors of $Q'(\omega)$ and $Q''(\omega)$, the Toeplitz inversion method discussed in Sec 5.6 is used. The computation of x_0'' using (7.4.1) is not preferred for the finite capacity case as the computation of $\tilde{\mu}_1''$ (as given by (7.3.40)) requires considerable computational effort. Since the dimension of $Q''(\omega)$ is not large, we have not used the recursive procedure. In view of the coupling between the two queues, the QLDs are iteratively computed. The iterative procedure is found to converge fast and typically it requires about 8 iterations for a relative accuracy of 10^{-12} . The QLDs at Q1 and Q2, obtained using both the exact model (model I) and approximate model (model II) are shown in Fig 7.3 and Fig 7.4 respectively. The QLDs at Q1 and Q2 are also evaluated using the simulation routine discussed in Sec 5.8 and the results are also presented in Fig 7.3 and Fig 7.4. The modifications required in the simulation routine to account for the finite capacity of Q2 have already been mentioned. The results corresponding to both computations and simulations for $N_1=150$ ($\rho = 0.35$) and $N_2 = 240$ ($\rho = 0.56$) are shown in Fig 7.5 and Fig 7.6.

In the next two examples we assume the capacities of Q1 and Q2 to be 475 and 25 respectively. In the first example, N_1 and N_2 are chosen to be 225 and 180 respectively. The traffic offered at Q1 and Q2 become (0.525, 0.42). The QLDs at Q1 and Q2 obtained using both computation and simulation are shown in Fig 7.7 and Fig 7.8. In the second example N_1 and N_2 are chosen to be 165 and 240 respectively. The traffic offered at Q1 and Q2 become (0.385, 0.560). The QLDs at Q1 and Q2 obtained using both computations and simulations are shown in Fig 7.9 and Fig 7.10. From these figures it can be concluded that the QLDs of Q1 and Q2, obtained using exact model agree well with the simulation results for both Q1 and Q2. The point estimates of the queue lengths as well as the busy period distribution are obtained by choosing long runs for the simulation. The number of cells served from both Q1 and Q2 are of the 10^8 in

each of these runs. Using the equations given in Sec 5.8, the confidence intervals were also computed. The 95% confidence interval was found to lie from 1 to 15 % of the point estimates in the entire range. Hence, the results obtained through the simulation may be considered to be sufficiently reliable.

The QLDs of Q1 obtained using the approximate model differ from the exact model results at higher queue lengths. The QLDs of Q2 obtained using the approximate model agree well with the simulation results in all these examples. In all these figures, we have also shown the traffic offered at Q1 and Q2 at each of the phases of MMPP 1 and MMPP 2. The phase transition rates of the MMPPs are also shown. From these values, it can be noted that when both the MMPPs are in phase 1 the traffic offered to the server becomes close to the capacity of the server and hence as noted in Sec 5.10 the approximate model under estimates the QLDs at Q1.

As the queue lengths at Q1 computed using model I and II differ only at the higher queue lengths, we expect that y'_0 and y''_0 will only be marginally different for models I and II. The probability of the low priority queue, Q1 being empty at an arbitrary time instant is computed for different values of N_1 for $N_2 = 180$ ($\rho'' = 0.42$) and $N_2 = 240$ ($\rho'' = 0.56$). (ρ'' is the average traffic offered at Q2). The results obtained using both model I and II are presented in Fig 7.11. From this figure it can be concluded that the results using both these models agree. The average queueing delays at Q2 are computed using the results of Chapter 6 for the both $\rho'' = 0.42$ and $\rho'' = 0.56$. The results obtained using model I, model II and simulation are presented in Fig 7.12 and are found to agree well.

REFERENCES

- 1 C Blondia , " The N/G/1 finite capacity queue ", Commun Statistic Stochastic Model, 5(2), pp 373-294 (1989)
- 2 M F Neuts, " Moment formulas for the Markov renewal branching process", Adv Appl Prob 8, pp 690-711, 1976
- 3 V Ramaswami, "The busy period of queues which have a matrix-geometric steady state probability vector", Opsearch 19, 1982, pp 238-261
- 4 Hunter, J J , "On the moments of Markov renewal processes", Adv Appl Prob 1, 1969, pp 188-210
- 5 H Heffes and D M Lucantoni," A Markov Modulated characterization of Packetized voice and Data traffic and Related Statistical Multiplexer Performance", IEEE J SAC , No 6, pp 856- 867, Sep 1986

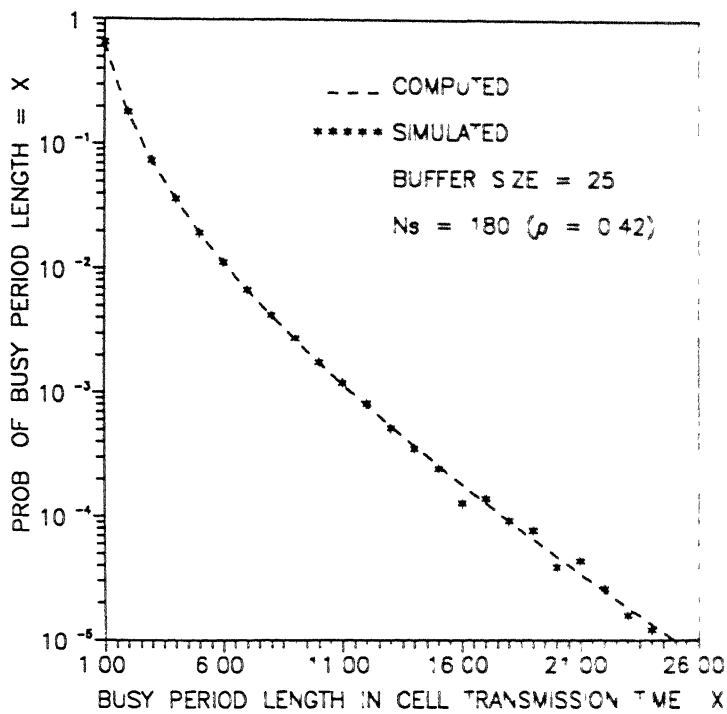


FIG 7.1 PMF OF THE BUSY PERIOD LENGTH OF MMPP/D/1/K QUEUE OBTAINED THROUGH COMPUTATION AND SIMULATION FOR $\rho = 0.42$

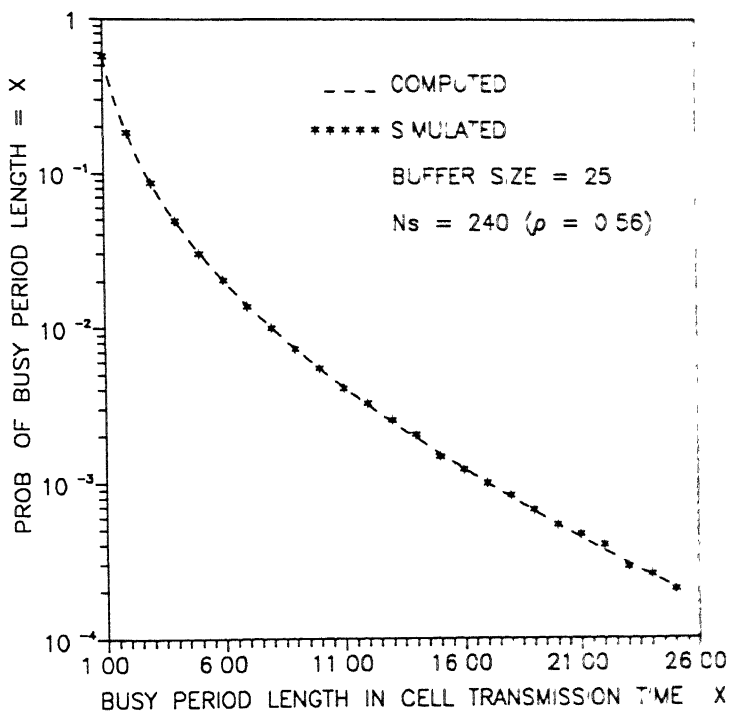


FIG 7.2 PMF OF THE BUSY PERIOD LENGTH OF MMPP/D/1/K QUEUE OBTAINED THROUGH COMPUTATION AND SIMULATION FOR $\rho = 0.56$

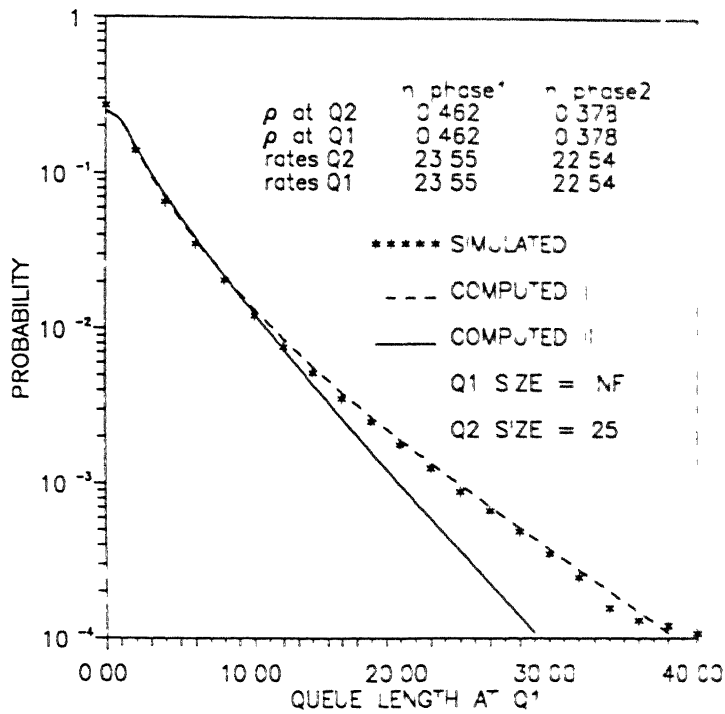


FIG 7.3 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED
 FOR (180,180) TYPE 1 SOURCES AND Q2 SIZE =25 ($\rho = 0.42, 0.42$)

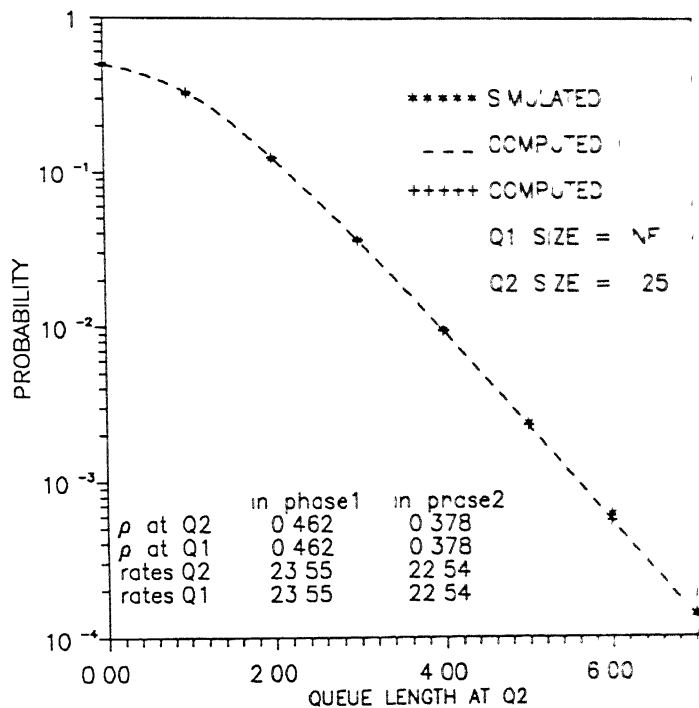


FIG 7.4 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED
 FOR (180,180) TYPE 1 SOURCES AND Q2 SIZE =25 ($\rho = 0.42, 0.42$)

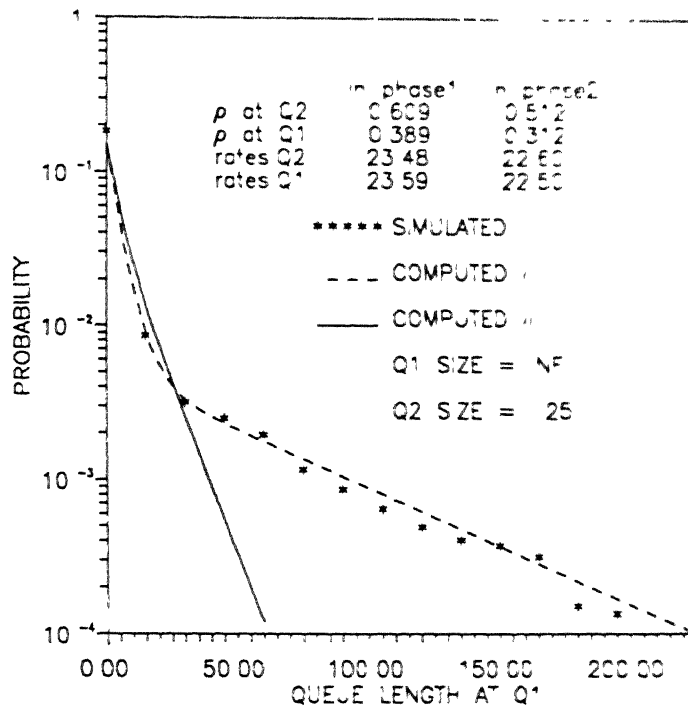


FIG 7.5 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED
 FOR (150,240) TYPE 1 SOURCES AND Q2 SIZE = 25 ($\rho = 0.35, 0.56$)

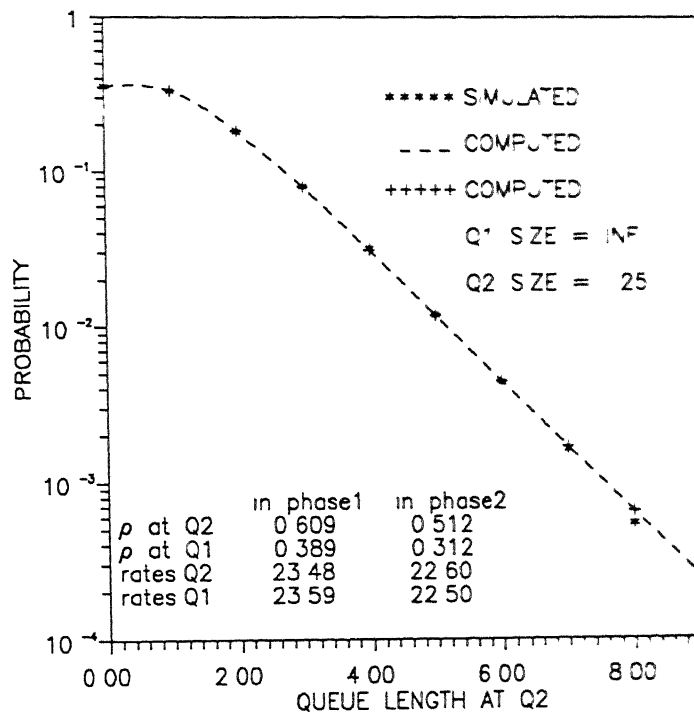


FIG 7.6 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED
 FOR (150,240) TYPE 1 SOURCES AND Q2 SIZE = 25 ($\rho = 0.35, 0.56$)

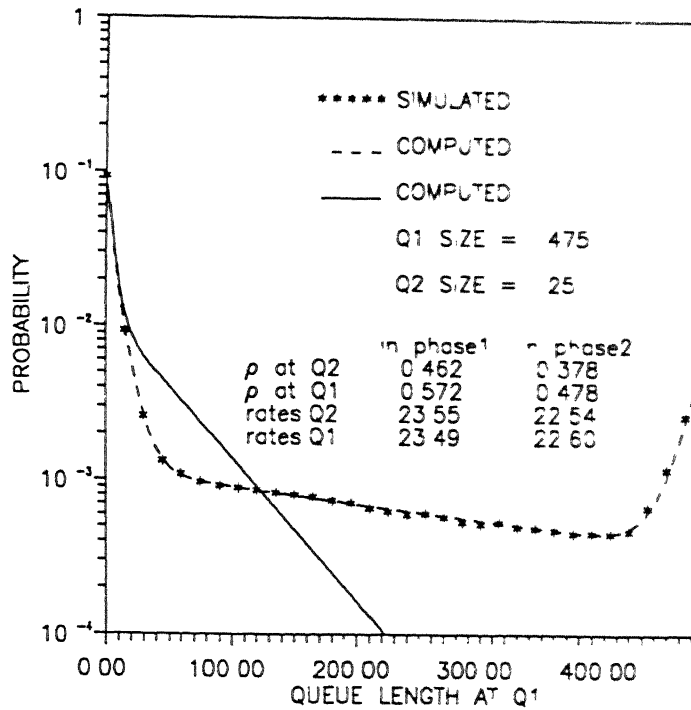


FIG 7.7 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED FOR (225,180) TYPE 1 SOURCES AND Q2 SIZE =25 ($\rho = 0.52, 0.42$)

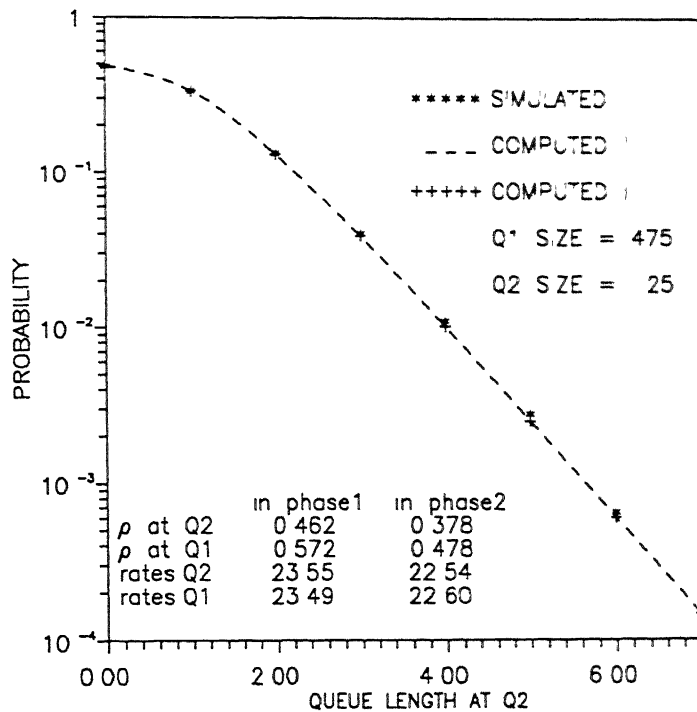


FIG 7.8 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED FOR (225,180) TYPE 1 SOURCES AND Q2 SIZE =25 ($\rho = 0.52, 0.42$)

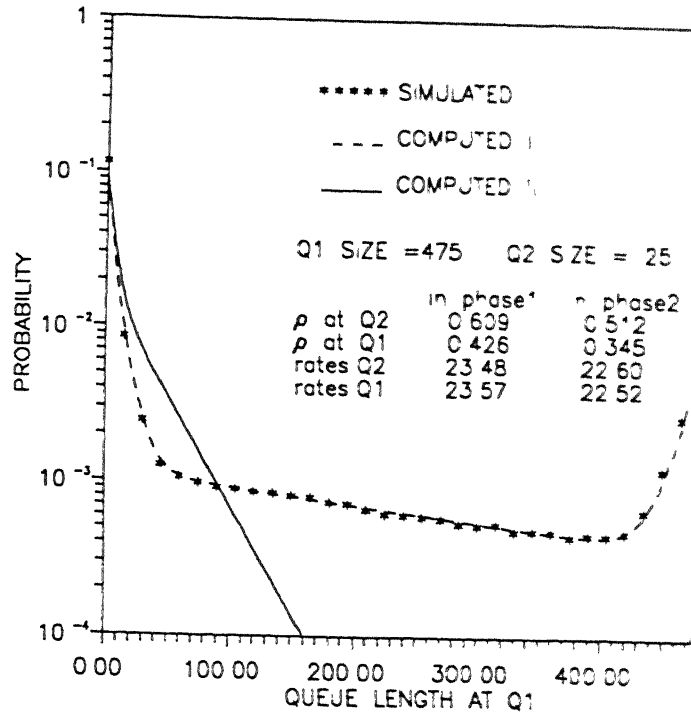


FIG 7.9 QLD OF Q1, THE LOW PRIORITY QUEUE COMPUTED AND SIMULATED
 FOR (165,240) TYPE 1 SOURCES AND Q2 SIZE = 25 ($\rho = 0.385, 0.56$)

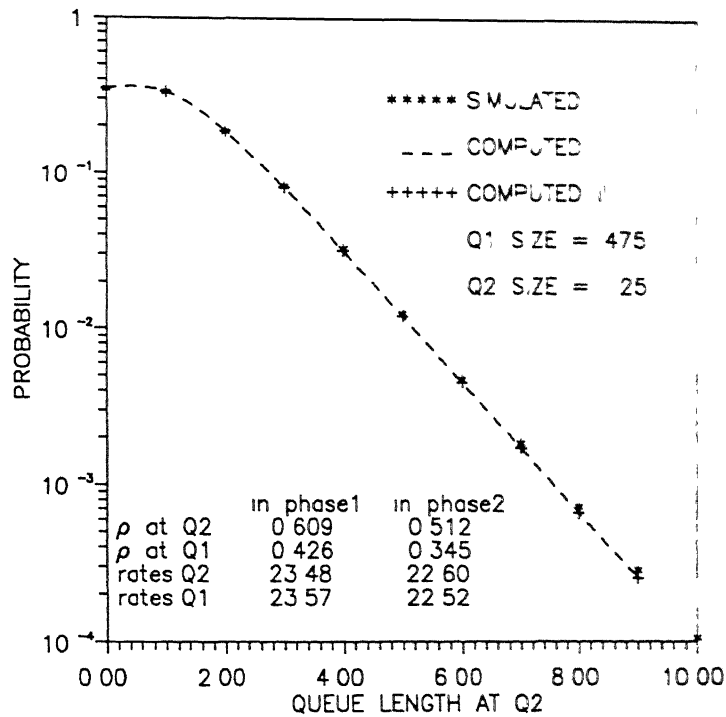


FIG 7.10 QLD OF Q2, THE HIGH PRIORITY QUEUE COMPUTED AND SIMULATED
 FOR (165,240) TYPE 1 SOURCES AND Q2 SIZE = 25 ($\rho = 0.385, 0.56$)

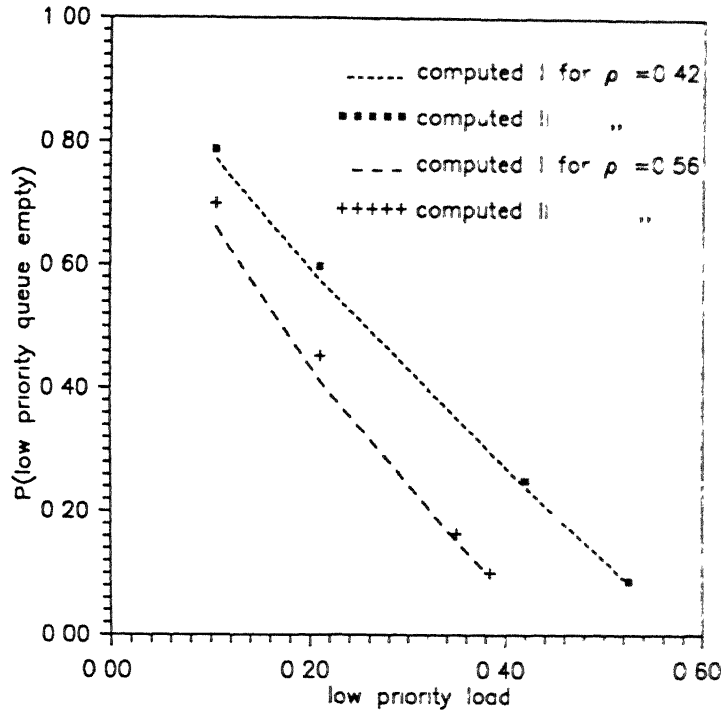


FIG 7.11 Variation of $P(\text{low priority queue empty})$ with traffic offered in Q1 and Q2 in an MMPP/D/1/K queue

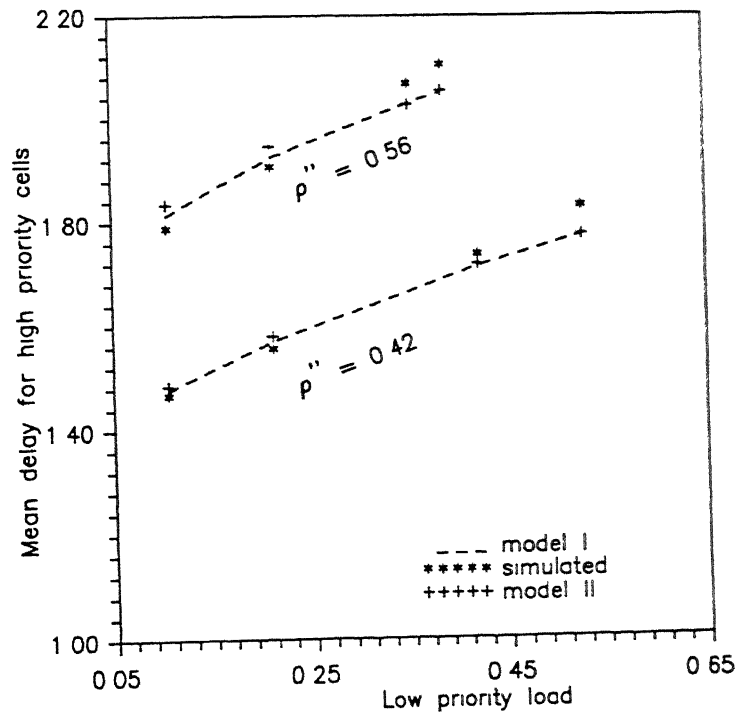


Fig 7.12 Mean delay for Q2 cells in an MMPP/D/1/K priority system obtained using model I, II and simulation

CHAPTER 8

EXTENSION FOR THREE OR MORE PRIORITY CLASSES

8.1. INTRODUCTION

In this chapter, We study a non-preemptive MMPP/D/1 priority system with more than two priority classes. As in the previous chapters, we assume the traffic from each priority class to arrive at separate queues and the service time is the same for each priority class ie D sec/customer. The queues are labelled in the increasing order of priority such that Q_1 is the lowest priority queue and Q_i is the highest priority queue in a system having i priority classes. We first consider the case where the number of priority classes is equal to three. For this the evaluation of the queue length densities at the departure instants of customers from the respective queues as well as the average queueing delays at these queues are considered. The extension of these results for the case with more priority classes is considered subsequently.

8.2 EVALUATION OF THE QLDs OF A NON-PREEMPTIVE MMPP/D/1 TRIPLE PRIORITY SYSTEM

In this section, we consider the case where the number of priority classes is equal to 3. The traffic from the three priority classes arrive at Q_1 , Q_2 and Q_3 and Q_3 has the highest priority. The QLD of Q_3 at the departure instants of customers from Q_3 can be obtained as follows. If a departure from Q_3 leaves the system non-empty, the next departure from this queue occurs exactly D sec later and its departure instant is independent of the state of Q_1 and Q_2 . If the departure leaves the system empty, the next departure instant from Q_3 does depend on the state of Q_1 and Q_2 . However, the effect of the customers in both Q_1 and Q_2 on the waiting time of the first customer arriving at the empty Q_3 is essentially the same. They delay his service for a maximum

of D sec Hence for the evaluation of the QLD at Q3, the customers at both Q1 and Q2 may be considered to belong to the same class and to arrive at a hypothetical queue denoted as Q12 Now the dual priority system with queues Q12 and Q3 with Q3 as the high priority queue can be solved using the steps given in chapter 5 This procedure, yields the QLD of Q3, the probability of Q3 being empty and the probability of either Q1 or Q2 being empty at an arbitrary time instant It may be noted that since we are interested only on the QLD of Q3, the QLD of Q12 need not be evaluated It may be recalled that we have indicated how the QLD of the high priority queue can be evaluated without evaluating the QLD of the low priority queue in Sec 5 6

The evaluation of the QLD of Q1 at the departure instant of customers from Q1 is considered next It may be noted that Q1, being the lowest priority queue, receives service only when both Q2 and Q3 are empty Given that the previous departure from Q1 left the system non-empty, the inter departure time of the next customer from Q1 consists of one customer service and the time required to serve all the customers that might have arrived at both Q2 and Q3 during this service time and the busy periods required by these customers Similarly, if the first customer arriving at an empty Q1 finds either Q2 or Q3 to be non-empty, it receives service only after both these queues become empty As far as Q1 is concerned, the effect of the customers in both Q2 and Q3 on the Q1 customers is the same Hence for the computation of the QLD of Q1, the customers arriving at Q2 and Q3 can be considered to arrive at a single queue Q23 and treated with high priority compared to Q1 The QLDs of the dual priority system (Q1, Q23) can be obtained using the results of chapter 5 In this case the QLDs of both Q1 and Q23 should be found This procedure also yields the probability of Q1 being empty and the probability of either Q2 or Q3 being empty at an arbitrary time instant

The evaluation of the QLD of Q2 is considered next. Since the procedure for this is involved, we shall consider the equations required for this purpose in detail. We denote the parameters corresponding to Q1 and Q2 by superscripts of (') and (") respectively and those of Q3 without any superscript. For the study of this queue we have to keep track of the phases of the MMPPs to both Q2 and Q3 simultaneously. Let the number of phases of MMPP 2 and MMPP 3 be M and N respectively. As in Chapter 2, we consider the composite phase process obtained by superposing the phase processes of MMPP 2 and MMPP 3. We consider the hypothetical MMPPs MMPP $\underline{2}$ and MMPP $\underline{3}$, the phases of which are equal to that of the composite phase process. The arrival rates of these MMPPs are also chosen as in Chapter 2. Let τ_n'' denote the n^{th} departure epoch of customers from Q2. Let (X_n'', J_n'') denote the number of customers in Q2 system and the phase of MMPP $\underline{2}$ at τ_n'' . Given that the previous departure from Q2 left the system non-empty, the time at which the next departure from this queue occurs depends on the state of Q3. If Q3 is empty, it occurs D sec later. Otherwise it occurs after an additional period of one busy period of Q3. The duration of the busy period of Q3 depends on the phase of the MMPP to Q3 at the previous departure instant from Q2. Hence it can be verified that $(X_n'', J_n'', \tau_{n+1}'' - \tau_n'')$ forms a semi-Markov sequence with the state space $[0,1, \dots, MN]$. As in Chapter 2, the transition probability matrix of this SMC denoted as $Q''(t)$ and can be expressed in terms of $MN \times MN$ matrix mass functions $A_m''(t)$ and $B_m''(t)$ whose $(i,j)^{\text{th}}$ elements are defined as follows

$$[A_m''(t)]_{ij} = P(\text{Given that a cell departed from Q2 at time 0, leaving at least one cell in Q2 and the arrival process MMPP } \underline{2} \text{ in phase } i, \text{ the next departure occurs at no later than time } t \text{ with MMPP } \underline{2} \text{ in phase } j, \text{ and in the intervening period there were } m \text{ arrivals})$$

$[B_m''(t)]_{ij} = P\{\text{Given that a cell departed from Q2 at time 0, leaving Q2 empty and the arrival process MMPP } \underline{2} \text{ in phase 1, the next departure occurs at no later than time } t \text{ with MMPP } \underline{2} \text{ in phase } j, \text{ and in the intervening period there were } n \text{ arrivals}\}$

As in Chapter 2, we define $MN \times MN$ matrices $P''(m, t)$, $H''(t)$ and $U_k''(t)$ whose $(i, j)^{\text{th}}$ elements are defined as follows

$$[P''(n, t)]_{ij} = P[N''(t)=n, \underline{j}(t)=j \mid N''(0)=0, \underline{j}(0)=1]$$

$N''(t)$ No of arrivals at Q2 in $(0, t]$ from MMPP $\underline{2}$

$$[H''(t)]_{ij} = P[(\tau_n'' - \tau_{n-1}'') \leq t, \mid \underline{j}_{n-1}'' = 1 \text{ and } X_{n-1}'' > 0] \delta_{ij}$$

$$[U_k''(t)]_{ij} = P[\text{Busy period of Q2 starts at or before time } t, k \text{ arrivals at Q2 in } (0, t], \underline{j}(t) = j \mid X''(0)=0, \underline{j}(0)=1]$$

$A_m''(t)$ and $B_m''(t)$ can then be expressed as follows

$$A_m''(t) = \int_0^t dH''(\sigma) P''(m, \sigma) \quad m \geq 0, t \geq 0 \quad (8.2.1)$$

$$B_m''(t) = \sum_{k=1}^{m+1} U_k''(t-D) P''(m-k+1, D) u(t-D) \quad (8.2.2)$$

The computation of $A_m''(t)$ proceeds along the same lines as the computation of $A_m'(t)$ of the dual priority system considered in Chapter 2. These details have therefore not been presented here.

Next we consider the computation of $U_k''(t)$. Let us consider the first customer arriving at an empty Q2 and denote him as F2. F2 receives service immediately only if both Q1 and Q2 are empty. We refer to this case as case 1. F2 waits for receiving service under the following three cases.

Case 2. If Q3 is receiving service, F2 waits for the on going service as well as the additional busy period (that Q3 might require after this service) to be over.

Case 3. If Q1 is receiving service, F2 waits for the on going service to be

over If Q3 is empty at this instant then F2 begins service

Case 4 If Q1 is receiving service, F2 waits for the on going service to be over At this instant if Q3 is found to be non-empty, then the busy period of Q3 starts and Q2 receives service only after this period is over

The introduction of three priority classes increases the complexity of estimating the probability of occurrence of these four cases In the dual priority system, the first customer arriving at the low priority queue waits for receiving service if the high priority queue is found to be non-empty at that instant Hence the probability of this event is equal to the probability of the high priority queue being non-empty However, in the triple priority system, the first customer arriving at Q2 may find Q3 to be non-empty under the following two conditions (i) Q3 is actually receiving service (ii) At the time when the ongoing service started, Q1 was non-empty but Q3 was empty When the service was in progress, one or more customers arrived at Q3 Hence the probability of occurrence of case 2 should be computed as the probability of F2 finding the server to be busy with a Q3 customer

Let us define p_{s1} to be the probability of the server being busy with the customers from Q1 p_{s3} can be evaluated along the same lines as in Conway [1] and is given by

$$p_{s3} = \frac{\tilde{\mu}}{\hat{\mu}} \quad (8.2.3)$$

where $\tilde{\mu}$, $\hat{\mu}$ are the mean busy period and mean busy cycle of Q3 respectively These parameters can be computed considering the dual priority system (Q1, Q3) and using the equations given in Sec 4.5 and 4.6 For both event 2 and 3 to occur, F2 should find the server to be serving a Q1 customer The probability p_{s1} , that the server is busy serving a Q1 customer is given by

$$p_{s1} = \frac{\mu^* D}{\hat{\mu}'} \quad (8.2.4)$$

where μ^* and $\hat{\mu}'$ are the number of customers served during the busy cycle and the duration of the busy cycle of Q1 respectively. It may be recalled that the busy period of the low priority queue is defined in Chapter 2 to be the time that elapses between the instant when the first customer begins service and the instant when the queue becomes empty again. The intervening period may consist of one or more busy periods of the high priority queue. That is why the form of the equations for p_{s1} and p_{s3} are different. p_{s1} can be computed by considering the dual priority system (Q1, Q23) and using the equations given in Sec. 4.3 and 4.4.

The quantity $\frac{d}{dt}U_k''(t)$ can be evaluated under the condition that case 1 is true for $i=1,2,3$ and 4 and the details for this are given in Appendix (8A). Multiplying these expressions by the probability that case 1 is true, we get

$$\begin{aligned} \frac{d}{dt}U_k''(t) &= (1-p_{s1}-p_{s3}) P''(0,t)\underline{\Lambda}'' \delta_{1k} \\ &+ p_{s3} \left\{ \frac{1}{D} \sum_{n=n_0}^{k-1} \sum_{\iota=0}^{\infty} \int_0^{\infty} Du(t-\iota+1D_-) + (t-\iota D)u(\iota+1D-t) \right. \\ &\quad \left. P''(0,t-\iota D-u)\underline{\Lambda}'' du P''(n,u)F(\iota)P''(k-n-1,\iota D) \right\} \\ &+ p_{s1} \left\{ \frac{1}{D} \int_0^{Du(t-D_-) + tu(D-t)} P''(0,t-D)P^*(0,0,D-u)\underline{\Lambda}'' du P^*(k-1,0,w) \right\} \\ &+ p_{s1} \left\{ \frac{1}{D} \sum_{n=1}^{\infty} \sum_{n=0}^{k-1} \sum_{\iota=1}^{\infty} \int_0^{\infty} Du(t-\iota+1D_-) + (t-\iota D)u(\iota+1D-t) \right. \\ &\quad \left. P''(0,t-\iota D-u)\underline{\Lambda}'' du P^*(n,n,u)P(n,k,n,\iota) \right\} \end{aligned} \quad (8.2.5)$$

The matrices $P^*(\cdot)$ and $P(\cdot)$ are defined in Appendix (8A).

Knowing the matrices $A_m''(t)$ and $U_k''(t)$, $Q''(\omega)$ can be found. The invariant probability vector of $Q''(\omega)$ gives the QLD of Q2. This completes the evaluation of the QLDs of a triple priority system.

8.3. EXTENSION FOR MORE THAN THREE PRIORITY CLASSES

First, we consider the application of the results obtained for evaluating the QLDs of a non-preemptive MMPP/D/1 priority system in which the number of priority classes are four. As in the previous section we assume the traffic from each priority class to arrive at separate queues and customer from each priority class demands the same service time of D sec/customer. The labelling of these queues are also same as before, i.e. Q_1 refers to the lowest priority queue.

The quadruple priority system can be split into two triple priority system. For example, the customers arriving at queues Q_1 and Q_2 can be treated to arrive at a single queue Q_{12} and considered to belong to a single class as far as the higher priority classes are concerned. Now this triple priority system consisting of queues Q_{12} , Q_3 and Q_4 can be studied using the results of Sec 8.2. This gives the QLDs of Q_3 and Q_4 in addition to that of Q_{12} . To obtain the QLDs of Q_1 and Q_2 , the traffic arriving at both Q_3 and Q_4 can be treated to arrive at a single queue Q_{34} and can be offered the highest priority. Now, the triple priority system consisting of the queues Q_1 , Q_2 and Q_{34} can be analysed using the results of Sec 8.2. This gives the QLDs of Q_1 and Q_2 in addition to that the QLD of Q_{34} . Thus the QLDs at all the four queues can be found.

The same ideas can be extended for any number of priority classes and the QLDs can be recursively computed. i.e. knowing the QLDs of a 4 priority system, the QLDs of a 5 priority system can be found. This in turn can be used to find the QLDs of a 6 priority system and so on. It should be mentioned here that the assumptions that the service/customer is constant and is the same for each priority class allows us to combine the various priority classes and study them using a lower priority system. In an ATM network, the cell size is the

same for customers with different requirements on QOS and hence customers with different priority levels do require the same service time/cell. Hence these assumptions are valid in an ATM network.

8.4. COMPUTATION OF THE AVERAGE QUEUEING DELAYS OF THE MULTI-PRIORITY SYSTEM

Computation of the average queueing delays of a non-preemptive MMPP/D/1 priority system with more than two priority classes is straight forward and requires less computational effort compared to the evaluation of the QLDs at these queues. Knowing the average queueing delays, the average queue lengths at the various queues can be obtained using Little's formula (see for e.g. Kleinrock [2]).

First we consider the triple priority system. As in Sec. 8.2, we can split the triple priority system into two dual priority systems (Q1, Q23) and (Q12, Q3). The average queueing delays of these dual priority systems can be obtained using the results of Chapter 6. This gives the average queueing delays at Q1 and Q3. The average queueing delay at Q2 can be found using the M/G/1 conservation law (see for e.g. Kleinrock [3]) as follows. Let us consider a composite queue denoted as Q to which the traffic from all the priority classes are fed and served on a First Come First Served (FCFS) basis. Let the average traffic offered and average queueing delay of the customers at this queue be ρ_T and W_T respectively. By the M/G/1 conservation law, given that the service disciplines conserve the work, the average queueing delay is independent of the service discipline. Hence the average queueing delay at Q is related to that of Q1, Q2 and Q3 as follows. Let the traffic offered and the average queueing delays at Q1, Q2 and Q3 be denoted as (ρ_1, ρ_2, ρ_3) and (W_1, W_2, W_3) respectively. Using these parameters W_T is given by-

$$\rho_T W_T = \rho_1 W_1 + \rho_2 W_2 + \rho_3 W_3 \quad (8.4.1)$$

Since the average queueing delays at Q, Q1 and Q3 are known, the average queueing delay at Q2 can be found

Extension of these results for the case where the number of priority classes are more than 3 is straight forward. For example, to compute the average queueing delays at the quadruple priority system, we consider the hypothetical triple priority systems (Q1, Q2, Q34) and (Q12, Q3, Q4). By obtaining the queueing delays of these triple priority systems, the queueing delays at all the priority queues can be found. The queueing delays of the higher priority systems can also be recursively computed in a similar fashion.

APPENDIX (8.A)

In this appendix, the details on the evaluation of $\frac{d}{dt}U_k''(t)$ under the condition that case 1 is true, (for $i=1,2,3$ and 4) are presented.

Case 1 In this case the busy period of Q2 starts with a single customer. If t is the time at which the BP of Q2 starts and at time 0 if Q2 was empty then

$$\left. \frac{d}{dt}U_k''(t) \right|_{t=t, \text{ case 1}} = P''(0,t)\underline{\Lambda}'' \delta_{1k} \quad (8.A.1)$$

Case 2 Let t be the time when the busy period of Q2 starts. As shown in Fig 8.1 the interval $(0,t)$ consists of three parts: an initial idle period when Q2 is empty, the residual service time (RST) of the Q3 customer and the Additional busy period of Q3 (ABP). Let the duration of these three intervals be $(t-\omega-\iota D)$, ω and ιD respectively. Let the number of arrivals at the second interval be n . The probability that k arrivals occur in $(0, t)$ can be computed by considering the mutually independent events that 0, n , $k-n-1$ arrivals occur in the above three intervals. Let the phase of MMPP \underline{Z} be i, ℓ, m and j at times 0, $t-\iota D-\omega$, $t-\iota D$ and t respectively as shown in Fig 8.1

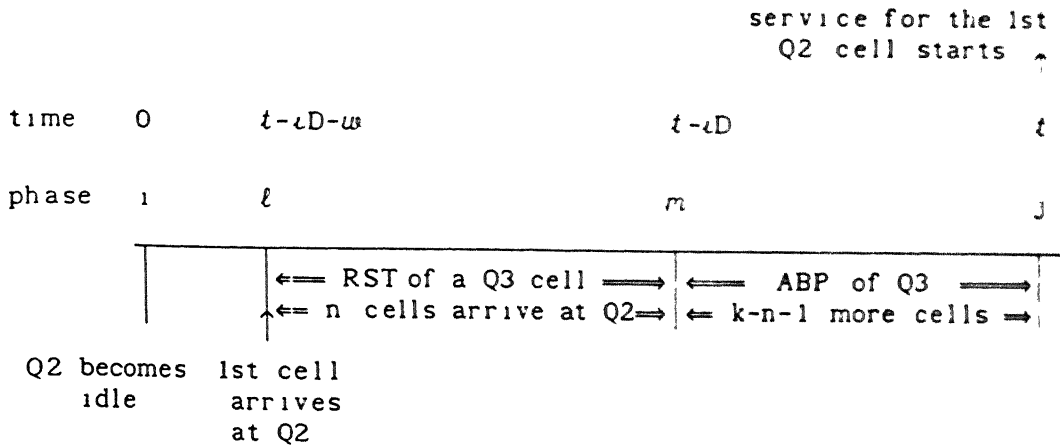


Fig 81 The various events that precede the BP of Q2 given that F2 finds the server to be busy with a Q3 customer

Considering the arrivals in the three intervals and keeping track of the phase of MMPP $\underline{2}$ we get

$$\left. \frac{d}{dt} U_{k,j}''(t) \right|_{\substack{\text{RST}=w \\ \text{ABP}=\epsilon D \\ t=t, \text{ case 1}}} = \sum_{n=n_0}^{k-1} \sum_{\ell=1}^N \sum_{m=1}^N P_{1\ell}''(0, t-\epsilon D-w) \Lambda_{\ell\ell}'' dw P_{\ell m}''(n, w) P_{mj}''(k-n-1, \epsilon D) \quad (8 A 2)$$

Here, n_0 denotes the minimum value of n . The value of n_0 depends on the existence of the ABP following the Residual service time of a Q3 cell seen by the first cell arriving at an empty Q1. When ABP is not present (i.e. $\epsilon=0$), then out of the k cells present at the time when BP of Q1 starts, $k-1$ cells arrive during the RST of the Q3 cell undergoing service. Hence the minimum value of n is equal to $k-1$ in this case. When ABP is non-zero, n_0 is zero and hence n_0 is given by

$$n_0 = (k-1)\delta_{\epsilon 0} \quad (8 A 3)$$

We also define the $MN \times MN$ diagonal matrices $F(k)$ whose i th diagonal element denotes $P[\text{ABP} = kD \text{ sec} \mid \text{ABP starts at time 0 with MMPP } \underline{2} \text{ in phase } i]$, i.e. the probability that an ABP of Q3 is of length kD sec given that it

started at time 0 with MMPP $\underline{2}$ in phase 1. Removing the condition on the length of ABP using the matrices $F(k)$, (8 A 2) becomes -

$$\left. \frac{d}{dt} U_{ij}^{(k)}(t) \right|_{\substack{RST=u, \\ t=t, \text{ case 1}}} = \sum_{n=n_0}^{k-1} \sum_{\ell=1}^N \sum_{m=1}^N P_{i\ell}^{(0,t-D-u)} \Lambda_{\ell\ell}^{(n,u)} \frac{d}{du} P_{\ell m}^{(n,u)} [F(u)]_{mm} P_{mj}^{(k-n-1,D)} \quad (8 A 4)$$

This can be written in matrix form as

$$\left. \frac{d}{dt} U_k^{(t)} \right|_{\substack{RST=u, \\ t=t, \text{ case 1}}} = \sum_{\ell=0}^{\infty} \sum_{n=n_0}^{k-1} P^{(0,t-D-u)} \Lambda^{(n,u)} \frac{d}{du} P^{(n,u)} F(u) P^{(k-n-1,D)} u(t-D) \quad (8 A 5)$$

The residual service time of a Q3 cell seen by the first cell arriving at an empty Q2 is also uniformly distributed between the interval (0,D) and hence its pdf is given by (3 3 4). Removing the condition on the RST using (3 3 4) and noting that when $0 < t-D < D$, the maximum RST is $t-D$, we get the terms inside square bracket of the second term of (8 2 5).

Case 3 Let t be the time when BP of Q2 starts. Let u be the residual service time of a Q1 customer. Then F2 arrives at time $t-u$ and $k-1$ more customers arrive at Q2 during the RST of Q1. For case 2 to be true, no arrival should occur at Q3 in the interval $(t-D, t)$. Hence we consider the following three intervals as shown in Fig 8 2: $(0, t-D)$, $(t-D, t-u)$ and $(t-u, t)$. In the last two intervals there should be no arrivals at Q3. The probability that k arrivals occur in $(0, t)$ can be computed by considering the mutually independent events that 0, 0, $k-1$ arrivals occur in the above three intervals. Let the phase of MMPP $\underline{2}$ be i, ℓ, j and j at times 0, $t-D, t-u$ and t respectively as shown in Fig 8 2. Considering the arrivals in the three intervals and keeping track of the phase of MMPP $\underline{2}$ we get

$$\left. \frac{d}{dt} U_{k,ij}(t) \right|_{\substack{\text{RST}=\omega \\ t=t, \text{ case 2}}} = \sum_{\ell=1}^{MN} \sum_{j=1}^{MN} P_{1\ell}''(0, t-D) P_{\ell j}''(0, D-\omega) P_{\ell j}(0, D-\omega) \Lambda_{jj}'' d\omega P_{jj}''(k-1, \omega) P_{jj}(0, \omega) \quad (8 A 6)$$

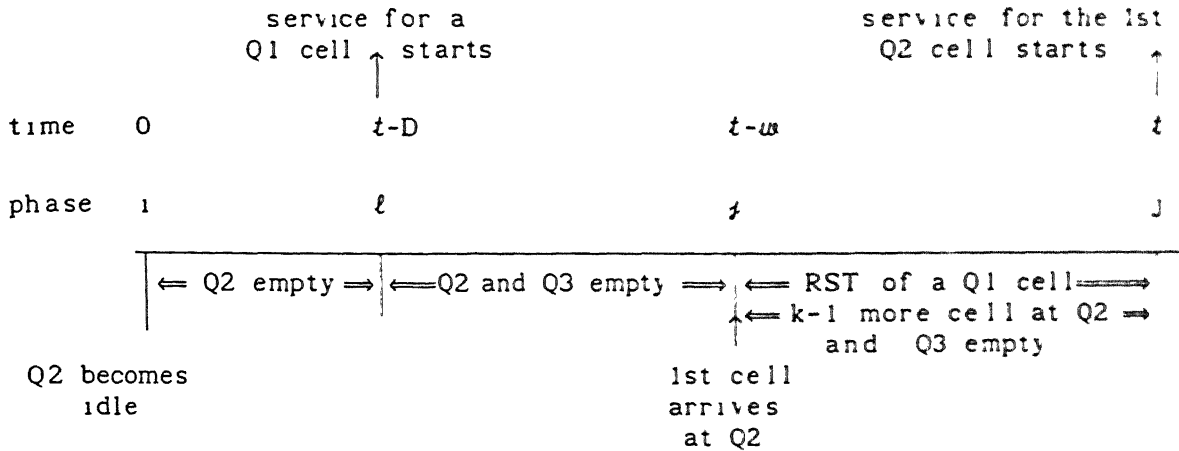


Fig 82 The various events that precedes the BP of Q2 given that F2 finds the server to be busy with Q1 and Q3 empty during the RST

To simplify the notation we define $MN \times MN$ matrices $P^*(n2, n3, t)$ whose $(i, j)^{th}$ are given by

$$\left[P^*(n2, n3, t) \right]_{i,j} = P_{1j}''(n2, t) P_{ij}(n3, t) \quad (8 A 7)$$

It may be recalled that the parameters of Q3 are written without the superscripts and the phase of MMPP 2 is equal to that of MMPP 3 at all time instants. Using (8 A 7) in (8 A 6), writing it in matrix form and removing the condition on the RST we get the terms inside the braces of the 3rd term of (8 2 5)

Case 4 In this case during the residual service time of the Q1 customer n , n ($n > 0$) customers arrive at Q2 and Q3 respectively. The busy period of Q3 follows this RST. Let N denote the number of customers of Q3 with which the BP of Q3 starts. Let the duration of the BP of Q3 be ωD ($\omega = 1, 2, \dots$). Let t be the time at which the BP of Q2 begins and let there be k arrivals in $(0, t)$

In the interval $(0, t - \epsilon D - w)$ Q2 is empty. As shown in Fig 8.3, let the phase of the MMPP $\underline{2}$ at times 0, $t - \epsilon D - w$, $t - \epsilon D$ and t be i , ℓ , j and J respectively. Considering the arrivals at Q2 and the phase transitions of MMPP $\underline{2}$ in the three intervals shown in Fig 8.3 and using equation (8 A 7) we get

$$\left. \frac{d}{dt} U''_{k_{1j}}(t) \right|_{\substack{\text{RST}=w \\ \text{BP}=\epsilon D, N=n \\ t=t, \text{ case 4}}} = \sum_{\ell=1}^{MN} \sum_{n=0}^{k-1} \sum_{j=1}^{MN} P''_{i\ell}(0, t - \epsilon D - w) \Lambda''_{\ell\ell} dw P''_{\ell j}(n, \epsilon D) P''_{jJ}(k-n-1, \epsilon D) \quad (8 A 8)$$

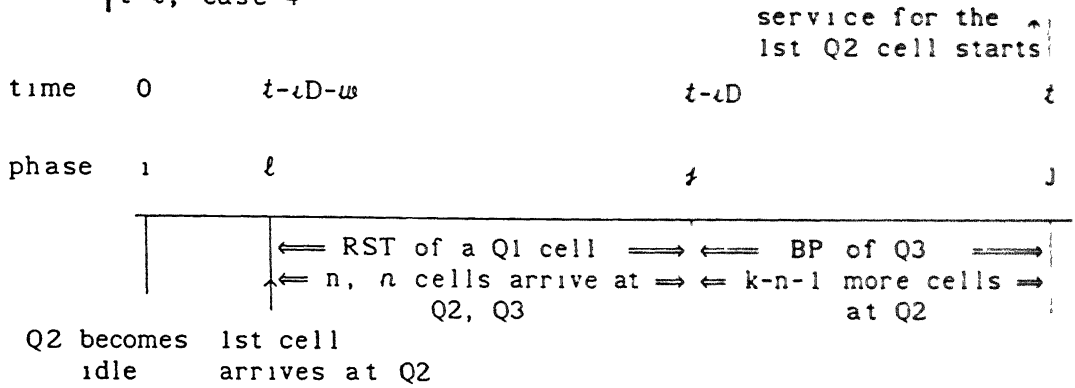


Fig 8.3. The various events that precedes the BP of Q2 given that F2

finds the server to be busy with Q1 first and then with Q3

To simplify (8 A 8), we define the $MN \times MN$ matrices G_{ℓ}^n whose $(i, j)^{\text{th}}$ element denotes the probability that the BP of Q3 is of length ϵD sec and ends with MMPP $\underline{2}$ in phase j given that it started at time 0 with n cells in Q3 and with MMPP $\underline{2}$ in phase i . Removing the condition on the length of BP using these matrices (8 A 8) becomes -

$$\left. \frac{d}{dt} U''_{k_{1j}}(t) \right|_{\substack{\text{RST}=w \\ t=t, N=n, \text{ case 4}}} = \sum_{\ell=1}^{MN} \sum_{n=0}^{k-1} \sum_{m=1}^{MN} P''_{i\ell}(0, t - \epsilon D - w) \Lambda''_{\ell\ell} dw P''_{\ell j}(n, \epsilon D) G_{\ell}^n P''_{jJ}(k-n-1, \epsilon D) \quad (8 A 9)$$

To simplify (8 A 9) further we define $MN \times MN$ matrices $P(n, \epsilon, k, n)$ whose $(i, j)^{\text{th}}$ elements are given by

$$\left[P(n, l, k, n) \right]_{i,j} = G_{ij}^n P_{ij}''(k-n-1, lD) \quad (8 A 10)$$

Using (8 A 10) in (8 A 9) and writing in it matrix form and removing the condition on the duration of the RST of Q1 as well as the condition on the number of customers with which the BP of Q3 starts we get the terms inside the braces of the last term of (8 2 5)

REFERENCES.

- 1 R W Conway, W L Maxwell and L W Miller, "Theory of scheduling", Addison-Wesley Publishing company, Massachusetts, 1967
- 2 L Kleinrock, "Queueing systems, Vol 1 Theory", Wiley, New York, 1975
- 3 L Kleinrock, "Communication Nets Stochastic message flow and delay", McGraw-Hill, New York, 1964

CHAPTER 9

CONCLUSIONS

9.1 INTRODUCTION

In this thesis, the queueing analysis of a non-preemptive MMPP/D/1/K priority system with either infinite or finite buffers, is carried out assuming the traffic from each priority class to arrive at separate queues (In a system with n priority classes we assume the traffic from each of these classes to arrive at n separate queues Q_i for $i=1,2,\dots,n$ with Q_n having the highest priority) The customers from each priority class are assumed to demand the same service time/customer. Computation of the queue length density (QLD) of the low and high priority queues (Q_1, Q_2) of the dual priority system is considered first. Computationally and storagewise efficient approximate models have also been considered. The computation of the average queueing delays at the low and high priority queues and the percentile of the queueing delay at the high priority queue are considered next. Finally, the extension of these results for the finite capacity system as well as for the case where the number of priority classes are more than two are considered. For both the infinite capacity and the finite capacity cases, the numerical results obtained using both the exact and approximate models are compared with those obtained using simulation for the dual priority system.

9.2 APPLICATION OF THE MATRIX ANALYTIC APPROACH FOR THE PRIORITY SYSTEM

For the queueing analysis, the matrix analytic approach is used. The study of the non-preemptive MMPP/D/1 priority system has been carried out in the frame work of "Queues of the M/G/1 type" with FCFS discipline by incorporating some generalization in the method used for the latter queues as given below.

- 1 The inter departure time of customers from the low priority queue depends on the phase of the arrival process to the higher priority queues and hence it should be treated as a vector random variable
- 2 In the priority system, the time when the first customer arrives at an empty queue and the time when the busy period of the server starts need not be identical. In view of this, the busy period of Q_i (for $i=1, 2$) is defined to be the time that elapses since the beginning of the service for the first customer arriving at Q_i and the time when Q_i becomes empty again

The above generalizations are also useful for the study of a queueing system in which the service time distribution depends on the state of the input arrival process

9.3 SOME OBSERVATIONS AND CONCLUSIONS DRAWN IN THE EVALUATION OF THE QLDs OF THE DUAL PRIORITY SYSTEM

The computation of the QLDs at each of the queues in a dual priority system is posed as the problem of evaluating the busy period distribution (BPD) of the higher priority queue, the counting functions associated with the MMPPs to Q_1 and Q_2 and the probability of finding the queues Q_1 , Q_2 empty at an arbitrary time instant. In order to arrive at an efficient procedure for the computation of the QLDs, the following problems have been considered

- 1 Development of efficient recursive procedures for the computation of the busy period distribution of the server in each of the queues when the buffer size is either infinite or finite. This approach does not require the inversion of the LST of BPD
- 2 Development of an efficient procedure for the computation of the counting functions associated with the MMPPs
- 3 Ensuring the convergence of the procedure for the computation of the above

counting functions at higher traffic rates

These problems are solved by exploiting the fact that the service time demand/customer is constant and is the same for each priority class

Numerical computation of the QLDs requires the evaluation of the invariant probability vectors of the transition probability matrices pertaining to Q1 and Q2. Towards this end the following issues are considered in detail

- 1 Computation of the probability of finding zero, one cell at the departure instant of cells from Q1 (for $i=1, 2$) using first passage time arguments
- 2 Computation of the probability of Q1 being empty at an arbitrary time t
- 3 Evaluation of the moments of the queue lengths at Q1 and Q2
- 4 Details and the relative advantages of the computation of the QLDs of Q1 and Q2 using (i) Gaussian elimination method (ii) Block Toeplitz inversion method and (iii) Recursive procedure

The following observations are made after considering the above issues

- 1 The evaluation of the QLDs using the recursive method requires the evaluation of x'_0 and x''_0 , or equivalently, the probability of Q1, Q2 being empty at their respective departure instants. The other two methods do not require this step
- 2 The evaluation of x'_0 and x''_0 involves several intermediate steps such as finding the average number of customers served during the busy periods of Q1 and Q2, finding the invariant probability vectors g' and g'' associated with the matrices $\tilde{G}'(z,s)$, $\tilde{G}''(z,s)$ characterizing the busy periods of Q1 and Q2
- 3 When the total traffic offered to the server is not very high, the Toeplitz matrix inversion method turns out to be more efficient than the other two methods for evaluating the QLDs. This is because this method exploits the structure of $Q'(\omega)$ and $Q''(\omega)$ for minimizing the computational efforts and

does not require the evaluation of x'_0 and x''_0

- 4 When the traffic offered to the server is close to the capacity of the server the computational and storage requirements of methods (i) and (ii) become prohibitively high and the recursive method becomes quite attractive
- 5 Of all the three methods, the Gaussian elimination method requires the minimum implementation effort

For the evaluation of the QLDs at high traffic rates, two alternate approaches are suggested. In one approach, the Q1 buffer size is treated to be finite and the resulting system is studied to evaluate the QLDs. In another approach, computation of the QLD of Q2 and the moments of the queue length at Q1 without evaluating the QLD of Q1 is considered.

It may be noted that the computational and storage complexity required for the evaluation of the QLD of Q1 and Q2 under the priority system is increased by a factor of $O(N^2)$ and $O(M^2)$ over that of the corresponding single priority system. (Here M, N denotes the number of phases of the MMPPs to Q1 and Q2). This is because in the priority system the phases of the MMPPs to both Q1 and Q2 need to be tracked at all departure instants. An approximate model which is computationally and storagewise efficient is proposed next. This model keeps track of the phase of only one of the MMPPs at a time. Using this model, the evaluation of the QLDs of Q1 and Q2 when the input to Q2 is approximated by Poisson processes, has also been carried out.

For validating the results obtained through the numerical computations, simulation routines have been developed. For the computation of the BPD of Q2 two simulation models have been used. In the first model, denoted as model A, the traffic to Q2 is assumed to be generated by N_2 identical on/off sources. In the second model denoted as model B, the traffic is assumed to be generated by a single MMPP source. The parameters of this source is assumed to be obtained

ned by superposing the N_2 on/off sources. For obtaining the QLDs of Q_1 and Q_2 , each queue is assumed to be fed by two independent MMPPs.

Results on the computation of the BPD of Q_2 and the QLDs of Q_1 and Q_2 using the exact model (model I) and approximate model (model II) are presented next for a number of examples and compared with those obtained using simulation. For numerical computations, We assume the traffic to Q_1 and Q_2 to originate from N_1 Type i on/off sources and N_2 Type j on/off sources, respectively ($i=j$ implies identical type of on/off sources to Q_1 and Q_2). An output link of 150 Mbps and a cell size of 53 bytes are assumed. The traffic to Q_1 and Q_2 are approximated by two 2 phase MMPPs. Knowing the MMPP model parameters, $Q'(\omega)$ and $Q''(\omega)$ are then found and the QLDs are obtained iteratively. Due to resource constraints, the evaluation of the QLD is considered only for cases where the traffic offered to Q_2 is less than or equal to 0.35. For cases where the high priority load is greater than 0.35 a finite capacity non-preemptive MMPP/D/1/K priority system is suggested for the evaluation. When the total traffic offered to the server is close to the capacity of the server the computational and storage requirements become high and hence in these cases Q_1 buffer size is assumed to be finite. Based on the examples considered the following conclusions are drawn:

1. The BPD of Q_2 obtained using simulation models A and B match well and the model B requires about 40% less computation time than that of model A.
2. The BPD of Q_2 obtained using the computation and simulation agree well.
3. The QLDs of Q_1 and Q_2 obtained using the exact model agrees well with those obtained using simulation in all the examples considered.
4. It appears that model II should not be used if either of the following two conditions are true:

(a) if λ_1''/λ_2'' is significantly larger than 1, where λ_1'' denotes the arrival

rate of the MMPP to Q2 in the i th phase and are labelled such that $\lambda_1^i > \lambda_2^i$

(b) there is a particular phase pair of the two MMPPs which tends to overload the server and this phase pair is fairly likely to arise

In these cases, the model I is recommended for computations. Otherwise, model II may be preferred due to its simplicity

- 5 The QLDs of Q1 computed using model II agrees with those of model I at low queue lengths even when conditions (a) and (b) are true. Because of this the probability of Q1 being empty at an arbitrary time instant computed using model I and II turns out to be essentially the same.
- 6 The QLD of Q2 computed using all the three methods agree under all conditions.
- 7 The iterative procedure for the evaluation of the QLDs of Q1 and Q2 converges fast. For a relative accuracy of 10^{-12} it requires about 8 iterations.

Finally, the results on the computation of the QLDs of Q1 and Q2 obtained by assuming the traffic to Q2 to be modelled as Poisson process, are presented for some typical examples. In this case, the QLD of Q1 agrees with that obtained using model II. The QLD of Q2 differs from those of model I and II at higher queue lengths.

9.4 EVALUATION OF THE QUEUEING DELAYS

The expressions for the distribution of the virtual waiting time of a customer arriving at Q2 and its Laplace Steiltjes Transform are obtained. Using these results, the average queueing delay at Q2 as well as that in Q1 are obtained. Extension of these results for the approximate model as well as the degenerate case of non-preemptive M/D/1 priority system are considered. For the examples considered earlier for the evaluation of the QLDs, the average queueing delays at Q1 and Q2 are computed using the exact as well as

approximate models and are found to be in agreement with the results obtained using simulation. For the M/D/1 system, the average queueing delays are computed and are found to be in agreement with the results obtained using an alternate approach. The expression for the LST of the virtual waiting time distribution also enables the computation of the percentile of the queueing delays at Q2.

9.5 FINITE CAPACITY DUAL PRIORITY SYSTEM

The computation of the QLDs of a non-preemptive MMPP/D/1 dual priority system with finite capacity at Q2 is also considered in this thesis. As in the infinite capacity case, the evaluation of the QLDs needs the computation of the BPD of Q2 with finite capacity. The BPD of the server in Q2 with finite capacity differs from the infinite capacity case as follows:

1. Customers arriving when the buffer is full are denied service.
2. The maximum number of customers that can be admitted into the system during the service time of a customer depends on the empty space in Q2. Because of this, the distribution of the first passage times are not identical and depends on the state of Q2.

As in the infinite buffer case, a recursive procedure for the computation of the BPD is developed, using the fact that the service time/customer is constant. Computation of the BPD for the finite capacity case requires considerable computational effort compared to that of the infinite capacity case. This is because, in the present case to compute the BPD of Q2 of capacity N , the BPD of capacity of 1, 2, ..., $N-1$ should be computed first. An indirect method for the efficient computation of the BPD of finite capacity Q2 is proposed. Modifications of the equations of the infinite capacity system for the present case is discussed.

The busy period distribution at Q2 is computed for some examples using both the direct and indirect methods. The latter method is found to require 50% less computation time. The BPD numerically computed is also compared with the simulation results and is found to match well. For some typical examples of traffic from on/off sources, the QLDs at Q1 and Q2 are evaluated using the exact model and the approximate model and compared with simulation results. The conclusions drawn for the infinite capacity is found to be valid for this case as well. Computation of the average queueing delays is also considered for the finite capacity case.

9.6 EXTENSION FOR THREE OR MORE PRIORITY CLASSES

Finally, the computation of the QLDs and the queueing delays of a non-preemptive MMPP/D/1 priority system with more than two priority classes are considered. The traffic from each priority class is assumed to arrive at separate queues and demand the same service time of D sec/customer. The computation of the QLD of a triple priority system is considered first and the extension of this result for higher number of priority classes are indicated. The computation of the average queueing delay also proceeds in a similar fashion.

9.7 SUGGESTIONS FOR FURTHER WORK

The following problems may be taken for further work:

- 1 The evaluation of the LST of the lower priority queues and the percentile of the queueing delays at these queues
- 2 Study of a non-preemptive priority system in which the service time/customer is constant but its actual value depends on the priority class
- 3 Study of service and space priority mechanisms together.



1899



Date Slip 800

This book is to be returned on the
date last stamped

--

EE-1994-D-VEN-QUE

